The Face *of* Text

Computer Assisted Text Analysis in the Humanities

Canadian Symposium on Text Analysis
November 19 – 21, 2004
McMaster University

Table of Contents

Introduction:

Program

Abstracts:

Message from McMaster University's Dean of Humanities

Welcome Researchers and Guests of the Face of the Text Conference! Many of you have come from abroad, Europe, the United States, and across the country. As Dean of the Faculty of Humanities, I extend warmest greetings to you all.

McMaster has a long tradition of supporting computing in the Humanities. Our first humanities computer lab was set up in the mid 1980s to support computer assisted language instruction. We have developed, and are justly proud of, an innovative and widely acclaimed Multimedia program. Our most recent initiative is the proposal for an interdisciplinary graduate program in Digital Society created by members of the Faculties of Humanities, Social Sciences and Engineering, which crosses faculty boundaries and will reach into the greater University community.

Research and learning in the area of digital arts and humanities is one of our Faculty priorities and we are delighted to host this third CaSTA (Canadian Symposium for Text Analysis). We are particularly pleased to be part of the TAPoR project that is developing innovative research infrastructure across Canada and here at McMaster. TAPoR and similar projects are transforming the way we deal with information on a global level. It is important that scholars such as yourselves lead in the adaptation of knowledge technologies to our various disciplines and continue to honour the Humanist tradition of innovation and creativity.

I welcome you to Hamilton and to McMaster and I wish you all a vigorous and successful conference.

Nasrin Rahimieh
Dean, Humanities

Welcome to The Face of Text

Welcome to **The Face of Text**, the third annual Canadian Symposium on Text Analysis, at McMaster University. We hope you will enjoy the three days of presentations and discussion.

This conference is organized by the TAPoR project. The Text Analysis Portal for Research is an internationally recognized research infrastructure project designed to support research using electronic media texts and computer assisted text analysis tools. This includes supporting research into text representation, text analysis tool development, text analysis techniques and theory, and the access and usability of electronic text environments. TAPoR has six contributing universities,

McMaster University. Contact: Geoffrey Rockwell - georock@mcmaster.ca and
Stéfan Sinclair sgs@mcmaster.ca
http://www.tapor.ca

University of Victoria. Contact: Raymond Siemens - siemens@uvic.ca
http://web.uvic.ca/hrd/tapor/

University of Alberta. Contact: Terry Butler - Terry.Butler@ualberta.ca
http://tapor.ualberta.ca/

University of Toronto. Contact: Ian Lancashire - ian@chass.utoronto.ca

Unversité de Montréal. Marc-André Morrisette - morissm@lexum.umontreal.ca
and Lexum - http://www.lexum.umontreal.ca/

University of new Brunswick. Contact: Alan Burk - burk@unb.ca
http://www.lib.unb.ca/TAPoR

At The Face of Text we will be demonstrating the first version of the TAPoR portal and inviting interested researchers to test the portal. If you want to learn more about the portal and get a test account e-mail Lian Yan (lyan@mcmaster.ca). For more information about TAPoR see: **http:www.tapor.ca.**

I would like to thank the people that made The Face of Text possible starting with the excellent speakers whose abstracts are included in these Proceedings. Thanks to the Social Science and Humanities Research Council of Canada (www.sshrc.ca) who funded the conference through their Image, Text, Sound and Technology program. McMaster University, OpenSky Solutions and Audcomp Computer Systems have also contributed generously. Paul Gallagher coordinated the conference. Andrew Paulin designed the web site (tapor1.mcmaster.ca/~faceoftext/), designed the Proceedings, and managed the publications. The program committee was composed of Terry Butler (chair), Elaine Toms, Peter Liddell and Stéfan Sinclair. Finally I would like thank the staff from Humanities Media and Computing for their support and all the students who are helping out.

Geoffrey Rockwell, McMaster University. TAPoR Project Leader and Conference Organizer.

Introduction to the Program

Terry Butler, Director, Research Computing, Arts Resource Centre, University of Alberta
Chair, Program Committee , CaSTA 2004

The process of inviting and then selecting papers for a scholarly conference is a specific example of a general rule: you had best be careful what you ask for; you might actually get it. This is vividly true in respect to the program for CaSTA 2004 – the Canadian Symposium on Text Analysis, which has emerged as an important scholarly venue for research and reflection on the nature of electronic text.

The CaSTA symposium was initially conceived as a research forum, which would coincide with the annual fall meeting of the TAPoR national executive. The first event, organized with great efficiency and on short timelines at the Université de Montréal, opened our eyes to the wide range of scholars working in the vineyard of text ("text", always, read broadly to include audio and video artifacts). The second CaSTA, at the University of Victoria, already had a strong national flavour, as well as highlighting the variety of advanced work being done at the host institution. The third, as the program clearly shows, is international both in its participation and the importance of the work being reported here.

We asked, in our Call for Papers: "how are electronic texts are made accessible to others, including non-specialists, students, and researchers in textual disciplines. How do text tools and electronic texts present themselves to users? What alternatives are there to search interfaces like Google?" We have got that, and much more. The papers in this program demonstrate best practices around interface, and they also contest the **idea** of interface; they show innovative modes of text visualization and representation, and cast doubt on the **validity** of representation.

There will also be ample opportunities for thoughtful reflection on the nature of the electronic enterprise. Beginning with Jerome McGann's opening keynote, the relationship between these forms of research and larger issues of academic life, and the world beyond, will be addressed. The era of uncritical "build it and they will come" is long past. We are in for a very interesting trip.

I am very pleased that McMaster University, our local hosts, have found a way to pack this substantial program into the short three days of the conference, and most importantly, to keep all the sessions in a single channel. As the cast of the papers makes abundantly clear, this research is arising from a vigourous conversation between and among traditional and emerging disciplines. It would be counter-productive to have the designers sitting in Room A while the technical types argue about code in Room B: it is precisely the intersection of these interests, and the exploitation of those intersections to ask questions, that has given rise to much of this research.

I would also like to thank my colleagues, Stéfan Sinclair (McMaster) and Peter Liddell (Victoria) who ably served on the program committee, and make the difficult task of winnowing our harvest to size both pleasant and stimulating. Also thanks to Christopher Poirier at Alberta, whose useful conference program review software made our task easier.

Program

Friday, November 19

8:00-9:00am	Registration
9:00-10:00am	Welcome Address
10:00-11:00am	Keynote Presentation: **Jerome McGann** The John Stewart Bryan University Professor, University of Virginia **Culture and Technology. The Way We Live Now; What is to be Done?**
11:00-11:30am	Break
11:30-1:00pm **Session 1:**	**David L. Hoover** English Department, New York University **(De)Facing the Text: Irradiated Textuality and Deformed Interpretations**
	Ray Siemens University of Victoria **Modelling Humanistic Activity in the Electronic Scholarly Edition**
	Patrick Juola & John Sofko Duquesne University **Proving and Improving Authorship Attribution Technologies**
1:00-2:00pm	Lunch
2:00-3:00pm	Keynote Presentation: **John Unsworth** Dean of the Graduate School of Library and Information Science, University of Illinois **Forms of Attention: Digital Humanities Beyond Representation**
3:00-3:30pm	Break
3:30-5:30pm **Session 2:**	**Gary W. Shawver** Humanities Computing Specialist, New York University **Oliver Kennedy** B.S. Computer Science, New York University **The Face of Meaning: A Short Presentation**
	Susan Brown University of Guelph **Facing the Deep: the Orlando Project Delivery System 1.0**

3:30-5:30pm
Session 2:

Michael Best
Coordinating Editor, Internet Shakespeare Editions, Department of English, University of Victoria
Show Me Your Image in Some Antique Book: Text and Image in the Internet Shakespeare

Mikaël Roussillon & Bradford G. Nickerson
Faculty of Computer Science, University of New Brunswick
Stephen Green & William A. Woods
Sun Microsystems Laboratories
Lightweight Morphology: A Methodology for Improving Text Search

Dominic Forrest
Université du Québec à Montréal
Automated text categorization: theory and application to computer-assisted text analysis in the humanities

Christian Vandendorpe
Département des lettres françaises, Université d'Ottawa
The Text of the database : can you read me ?

Evening
Meeting For TAPoR Research Board

Saturday, November 20

8:30-9:30am
Keynote Presentation: **Julia Flanders**
Brown University, Women Writers Project
Text analysis and the problem of pedantry

9:30-10:00am
Break

10:00-11:30am
Session 3:

Claire Warwick
School of Library, Archive and Information Studies, University College
Whose funeral? A case study of computational methods and reasons for their use or neglect in English literature

Paul A. Fortier
Centre on Aging / Department of French, Spanish and Italian, University of Manitoba
Computer Text Analysis and the Man from Wausau

Paul Scifleet & Concepcion S. Wilson
School of Information Systems, Technology & Management, The University of New South Wales
The Markup Analysis Engine: A new role for content analysis in digital content development

11:30-12:30	Lunch
12:30-1:30pm	Keynote Presentation: **John Bradley** Senior Analyst, Humanities Computing, King’s College London **What you (fore)see is what you get: Thinking about usage paradigms for computer assisted text analysis**
1:30-2:00pm	Break
2:00-3:30pm	McMaster Presentations: 1. **Jenna Wells & Madeleine Jeay** McMaster University **Testing TAPoR with Medieval French Poetry** 2. **William Coleman & Andrew Mactavish** McMaster University **Disseminating Research in Accessible Ways: the *Globalization and Autonomy Compendium*** 3. **Nicholas Griffin** McMaster University **James Chartrand** Open Sky Solutions **The Collected Letters of Bertrand Russell: An Electronic Edition** 4. **Stéfan Sinclair** McMaster University **HyperPo: A Longitudinal Perspective**
3:30-4:00pm	Break
4:00-5:30pm **Session 4:**	1. **Atefeh Farzindar & Guy Lapalme** RALI, Département d’Informatique et recherche opérationnelle, Université de Montréal **LetSum, a Text Summarization system in Law field** 2. **Tobias Kalledat** School of Business and Economics of the Humboldt University in Berlin **Automatic Trend Detection and Visualization using the Trend Mining Framework (TMF)** 3. **Stan Ruecker** Assistant Professor, Humanities Computing, University of Alberta **Zachary Devereux** MA Candidate, Department of Political Science, University of Alberta **Scraping Google and Blogstreet for Just-in-Time Text Analysis**
6:00-9:00pm	Reception and Banquet Dinner

Sunday, November 21

8:00-9:00am	Coffee
9:00-10:00am	Keynote Presentation: **Jean-Guy Meunier** Professor, Université de Québec à Montréal **Interfacing the text: difficulties and solutions**
10:00-10:30am	Break
10:30-12:30pm **Session 5:**	1. **Pamela Asquith** Professor, Department of Anthropology, University of Alberta **Peter Ryan** Doctoral student in the Department of Communication and Culture, Ryerson University **From System-Centered to User-Centered Design: The Kinji Imanishi Digital Archive Project**
	2. **Eugene W. Lyman** Boston University **In pursuit of radiance: Report on an interface developed for the Piers Plowman Electronic Archive**
	3. **Marc Plamondon** University of Toronto **Computer-assisted phonetic analysis of English poetry**
	4. **Jason Boyd** Ph.D. Candidate, Department of English, University of Toronto **REED Unbound: Accessibility, Interpretation, and the Patrons and Performances Web Site**
	5. **Elaine Toms** Dalhousie University **Geoffrey Rockwell & Stéfan Sinclair** McMaster University **Ray Siemens** University of Victoria **Modelling the humanities scholar at work**
12:30-1:30pm	Keynote Presentation: **Stephen Ramsay** Assistant Professor, University of Georgia **In Praise of Pattern**
2:30-3:00pm	Closing Remarks

From System-Centered to User-Centered Design: The Kinji Imanishi Digital Archive Project

Pamela Asquith, Professor, Department of Anthropology, University of Alberta
Peter Ryan, Doctoral student in the Department of Communication and Culture, Ryerson University

This archive fond is made up of 8000 pages of the personal notes and papers dating from 1919-1980 of Japanese scientist, anthropologist and explorer Kinji Imanishi (1902-1992). The collection is in Japanese, English and German and contains a wide array of paper types (from notebooks to air-letter paper) and sizes (from receipts to table-size maps), which are in a variety of conditions (like-new to almost disintegrating). The project began in 2002 with digital photography of a selection of the materials in Japan and continued in 2003 when we received permission for the entire collection to be brought to the University of Alberta for completion of the digital imaging and creation of the database for the images. In 2004 photoshop work was completed to prepare the images for the website that was concurrently developed.

Creating a balance between "user-centered" and "system-centered" (Cooper, *The Inmates are Running the Asylum,* 1997) technological solutions for the presentation of digital materials is a concern for any major digital project, especially given the wide array of technological solutions and tools described in contemporary academic journals such as Peter Flynn's "Is There Life Beyond the Web?" (*Journal of Literary and Linguistic Computing* 17.1, 2002); Mark Fraser's "From Concordances to Subject Portals: Supporting the Text-Centred Humanities Community" (*Computers and the Humanities* 34.3; August, 2000); Margaret Hedstrom's "Digital Preservation: A Time Bomb for Digital Libraries" (*Computers and the Humanities* 31.3, 1998) and Susan Schreibman's "Computer-mediated Texts and Textuality: Theory and Practice" (*Computers and the Humanities* 36.3; August 2002).

The two presenters provide perspectives on scholars' use of digital textual resources. Specifically, P. Asquith will speak on factors that affect a scholar's choice of what is of value to digitize, cross-cultural and multidisciplinary accessibility issues, and some of the benefits and barriers to effective use of digital textual resources that have become apparent thus far in the project. P. Ryan will speak on the project management of structuring the web site and database. For this archive, we found concrete solutions for our context-dependent design issues in the developing research areas of "interaction design" (Cooper, 1997) and "situated activity" (Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*, 1987).

Content Analysis of Short, Structured Texts: International Treaties

Harold W. Bashor, American Graduate School of International Relations & Diplomacy

How can one analyze short, structured texts? Content analysis provides a systematic, replicable technique where validation normally takes the form of triangulation. Such mutual convergence lends credibility to the findings by incorporating multiple sources of data, methods, investigators, or theories. However, this paper examines the strengths and weaknesses of triangulation by comparing data collection methods for long, unstructured texts with short, structured texts.

Considering the limitations of traditional triangulation, a multifaceted strategy has been tailored to meet the needs of specific, micro-linguistic analysis of a particular genre of text: international treaties. In addition, a number of different perspectives to discern the following key elements and the dynamics of text have been examined: (1) purpose or overall agenda, (2) agent and his/her social-cultural standpoint, (3) the intended audience, (4) the socio-political context, and (5) most relevant to this paper, the language itself: syntax, semantics, style, rhetoric, and structure.

The combination of these elements requires a multifaceted strategy that conventional triangulation often ignores in content analysis. This paper proposes a multi-layered approach with three different software programs. Although it has been argued in this paper that mutual validation of results is the ultimate goal, the capacity for methods to complement each other in the drive towards 'comprehensiveness' should not be assumed; nor should corroboration between methods be automatically considered unproblematic, and therefore precluding the need for further investigation.

Show Me Your Image in Some Antique Book: Text and Image in the Internet Shakespeare

Michael Best, Coordinating Editor, Internet Shakespeare Editions, Department of English, University of Victoria

In this paper, I shall discuss some of the ways in which images, both of early editions of the plays and of performances of them, will be made searchable in the Internet Shakespeare Editions.

As a "report of progress" in the development of a database of performance materials, the paper will look at ways in which image can be integrated with text, both by the editors of the plays and independently by visitors to the site.

REED Unbound: Accessibility, Interpretation, and the Patrons and Performances Web Site

Jason Boyd, Ph.D. Candidate, Department of English, University of Toronto

The aim of Records of Early English Drama (REED) is to find, transcribe, and publish external evidence of dramatic, ceremonial, and minstrel activity in Great Britain before 1642. REED volumes have been crucial for the historical grounding of scholarship concerning such activity, "transforming what is known about social and cultural life in Britain" in the period, in the words of one scholar. To date, the project has published 21 city and county collections comprising 23 volumes, ranging in size from the 262 page Newcastle upon Tyne to the 3-volume 1885 page Kent: Diocese of Canterbury. Given the liberal scope of REED's stated aim, and with a majority of England's counties, as well as Wales, Scotland, and London still to come, the sight of an ever expanding number of 'bright red volumes marching along the shelf' (to paraphrase one scholar) threatens to daunt even the most doughty user of the REED collection. The unwieldy staticness of the printed-volume format for researchers that wish to avail themselves of REED's entire corpus of published material becomes ever more pronounced as new volumes are produced. Unless one's research is focused, like the collections themselves, on an individual city or county, to locate references to, for instance, patronized musicians in the 16th century, or the role of women in entertainments of the gentry and nobility, requires time-consuming index-searching and analysis of all the relevant records in each collection, simply to compile the primary material.

The aim of REED's online database, the Patrons and Performances Web Site, is to facilitate this research by undertaking the challenging task of abstracting the 'hard' data from these often ambiguous and imperfect historical documents and by enabling the user to effectively search this data through multiple avenues and angles which encompass a spectrum of research interests. This web site will be essential in helping scholars to make effective use of what is not only a complex and continually expanding corpus of historical material, but it allows (in a way not possible in print format) for a mobility and manipulability of data that is responsive to the individual user's requirements.

More significantly, the web site allows for an enhancement of this frequently sparse primary data via its connection to a wealth of secondary, multimedia material. One can bring up a chronological list of all the recorded performances of a patronized troupe, learn where their patron resided and exercised influence through important offices and land holdings, and to which other patrons he or she was related, find out how a troupe was compensated for performances; learn about the architectural changes extant performance venues have undergone and the layout of possible performance spaces (via contemporary photographs, historical pictures, excavation schematics and antiquarian descriptions), and see what other troupes are recorded as having performed at the same venue. With interactive maps, one can see other proximate venues, the routes which traveling companies used and the venues they passed and possibly performed at in their travels. As a tool to enhance the REED volumes as a whole, this secondary material and its presentation is crucial to the web site's goal of making REED's scholarship accessible not only to a wider university audience, but to interested persons and groups in the broader community.

This combined paper/web site demonstration will discuss the challenges in transforming variegated primary documents into manipulable data, and illustrate how the web site format enables the user to easily and effectively access the full and complex range of primary data and secondary multimedia material.

What you (fore)see is what you get: Thinking about usage paradigms for computer assisted text analysis

John Bradley, Senior Analyst, Humanities Computing, King’s College London

Computer assisted text analysis has been one of the perennial topics in humanities computing, and has provided one of the mainsprings that has driven the field. Nonetheless, the results - particularly when viewed from the humanities more generally - have been disappointing. Tools that have been developed have met with few successes, and have had little uptake by the humanities community at large. As Yaacov Choueka said many years ago "the tools are here, where are the results?"

My talk will take some of its inspiration from the work and influence of Douglas Englebart, who in the late 1960s showed a developing computing community a fundamentally different way to think about computing and how it could fit into a user community. At the time, Englebart's thinking seemed to many to be wildly "out of the box". Nevertheless, his ideas became several of the driving forces that moved computing from being a minor interest in business, to being a tool used by virtually everyone in both the business and academia. My presentation is intended to encourage in us all similar "out of the box" thinking about computers and text analysis to see if what seems like a more radical approach can lead to a paradigm that will find a truly useful and broadly accepted role for computing in this fundamental humanistic activity.

Facing the Deep: The Orlando Project Delivery System 1.0

Susan Brown, the Orlando Project, University of Guelph

This paper describes the interplay between DTD design and subsequent developments that shaped the first full version of the Orlando Project’s delivery system. The project is currently completing a large-scale, intensively encoded body of digitally original materials on women’s writing in the British Isles from the beginnings to beyond the 1960s. Text markup generally proceeds either from the desire to represent a pre-existing text and thus an extant spatial arrangement of text on a page, or from a relatively clear sense of what effects the structural markup and any content-related markup will be used to produce. Neither was the case with the Orlando Project. The project’s extensive content-oriented tagset was developed to reflect its researchers’ priorities in literary history before any material had been written, and without any specifications for delivery beyond a basic allegiance to the TEI for representing document structure. The project DTDs therefore represented a set of abstract priorities for producing a literary history, untrammeled by specific ideas about delivery. The project's XML delivery system was developed years later.

To begin with a quick overview, the Project was conceived of as an experiment in interpretive markup more than a decade ago, when visual browsers were in their infancy and it was an open question whether Standard Markup Language in general and the Text Encoding Initiative in particular would be established as the standards they have become for the digital representation of texts in the humanities. The core literary team of project director Patricia Clements, Isobel Grundy, and myself were new to humanities computing and, taking advice from co-investigator Susan Hockey about the future direction of such work, jumped headlong in to the business of DTD development without any extensive experience of working with SGML, and with quite distinct aims from those of existing text encoding projects. These were to conduct an experiment in meeting some of the challenges that had been leveled at conventional literary history, by exploiting the capaciousness, flexibility, non-linearity, multivoicedness, and pressure towards self-reflexivity associated with electronic textuality. We wanted the electronic scholarship produced by the Orlando Project to use SGML both to provide the familiar visual organization of the text that our readers or users would expect, and to provide a new kind of self-reflexive structure that would conduct literary history in new ways.

So our design of the Document Type Definitions for the Orlando Project faced a number of challenges. Existing SGML projects and the TEI helped us decide on basic structural hierarchies for our documents, but took us only a certain distance as regards the content markup. We had no documents on which to perform document analysis. And we had no clear sense of how delivery would work. What we did instead was to try to define factors important to our sense of the history women's writing in the British Isles. Then we had to relate these concepts to each other in a hierarchical fashion, and reconcile the structural hierarchy, which we had adopted from the TEI to the hierarchy of the conceptual tags. This process of development, testing, and revision took intensive collaborative work over several years. Throughout this process, although we would occasionally remark, "Wouldn't it be cool if . . . ," we tended to bracket most questions of delivery, since the technical ground was shifting under our feet as we worked. It was by no means clear whether XML would be adopted by the Web and enabled by mainstream Web browsers.

The project devised 5 DTDs. Simple ones for chronology items and one for bibliographical information were conceived with the clearest sense of delivery: the granularity of these materials made them obvious candidates for dynamic manipulation by users to produce custom bibliographies and chronologies. The heart of the project is in its two DTDS for encoding discussions of lives and those of writings and literary careers. These we conceived as accounts in continuous prose, along the lines of short encyclopedia entries. We conceived these as somewhat dynamic, particularly insofar as they would embed chronology events that would become part of the project's chronology as a whole. And we also devised a fairly simple DTD for encoding discussions of topics not related to individual writers.

Now to give provide some idea of the design of the Life and Writing DTDs: the Life DTD uses Div1 structures to organize biographical discussion into major areas including Birth, Cultural Formation, Death, Education, Occupation, Names, Political Affiliation, and so on. These frequently have unique subtags and /or attributes, such as Birth position, Religion, and denomination. The Writing DTD contains many more unique tags than the Life one and thus appears more complex, particularly as these are organized into three subcategories of production, reception and textual features. However, it is in fact less restrictive than the biography DTD, making heavy use of inclusions to allow all the subtags from any of these three organizing categories to be invoked within a Div2 devoted to another one. This decision to create what is in effect a very flat DTD, which runs counter to the SGML desideratum of an ordered hierarchy of content objects, was the result of the sense that the imbrication of various issues in critical discussions was even greater than in biographical ones, and that to permit readable prose we had to allow for what were in essence overlapping hierarchies.

Eventually, we turned seriously to the development of a delivery system for the body of encoded texts we had created. Some early ideas, such as marrying our tagging of place with a geographical information system, remain unrealized for want of resources. But we are now testing the first full version of our delivery system.

The degree of dynamism within the system is considerably greater than initially conceived. Document pairs tagged with the Life and Writing DTDs are grouped with their bibliographic and any extrinsic chronological material to make up what we refer to as “entries”; these are available in the form of sets of tabbed screens and in that form are relatively static. However, they are also sectioned and dispersed by the system in several more dynamic ways. In addition to the chronological and bibliographical features already mentioned, these include:

Browsing: the system selects daily and provides as part of the home page a random set of hyperlinks offering serendipitous paths into project materials. These include dates, personal names, titles of text, place names, and organization names, and lead either to a screen of chronological items or the entry for an individual writer.

Links screens: automated hyperlinks are created for all multiple instances of personal names, places, organizations, and titles. The links are organized according to their occurrence within the major Life or Writing content tags. A user can thus move through the project materials in an informed and selective way. The links screens also make available a timeline and a set of excerpts, so readers can chose to pursue the hyperlink in the context of the source discussions, in chronological order, or in relation to other instances of the link.

Tag search results: users can search on tag occurrences or tag contents according to their particular interests. Results are returned as excerpts from entries, which can be read in a set (and in dialogue with their immediate tagging context, which is also provided), or can be used to access the full discussion in the source entries.

People entry point searches: searches for people writing in a particular genre or belonging to a particular occupational group result in both a list of names, for fuller exploration within the source entries, and a set of excerpts that can be read independently. This makes some of the functionality of the tag searches available with a less daunting interface.

All of these functions, and particularly the links and tag search results, result in a series of excerpts that at first glance might seem to resemble the scattershot outcome of an internet search. They are, however, far more than that, emerging as they do from a body of materials produced according to consistent research protocols and returned not simply on the basis of semantic content (although a free text search is available) but on the basis of the tagging structure. Thus, although there is no linking narrative, and indeed the set of excerpts may in some cases seem quite heterogeneous and even contradictory, this is, we would argue, a new form of literary history embedded in and produced by the encoding and its interaction with semantic content. For, as Jerome McGann argues of the Collex toolset being developed by the Applied Research in 'Patacriticism initiative, the very act of placing materials in relationship to one another, remediating them and juxtaposing them, makes an argument. In this case, the argument is formed in the interaction between the tagging structure devised and applied by the Orlando project team members who have produced the materials, and the user who brings her own preoccupations and shapes her inquiries of and pathways through the materials.

The challenge of allowing our users to plumb the project's unusual form of knowledge representation in an informed manner, which we have been aware of from the outset, has been emphasized by the results of our first round of user testing on the system. It became clear that we needed to declutter our search screens as much as possible, so as to clarify search logic. The vast majority of our testers suggested that we needed to resist trying to elucidate all the nuances of the system on the search panels. There was also a strong desire for more help and documentation, so perhaps this means that users will seek details and nuance after they have reached a basic level of familiarity.

As a consequence, while streamlining the screens, we have instituted help at a range of levels. The home page provides links to introductory screens with tips for getting started as well as very basic introductory searches. How Orlando Works provides a fuller introduction. Each entry point has a general help page explaining how it works, and a right-click help feature elucidates specific screen elements. Together, we hope these features will help users navigate the Orlando materials effectively.

Navigation is a major concern. Among our strategies to orient users and simplify navigation are: the hierarchical organization of the site into the home page and three major entry points; the use of tabbed screens within entry points and the hyperlinking screens to emphasis the relationships between sets of materials; and keeping screen proliferation to a minimum. Our user testing indicates that being able to through materials effectively is crucial: none of our testers reported having read right through individual entries on writers. This suggests that granularity will be a major criterion for the production of scholarship in electronic form. Thankfully, although the Orlando Project was conceived in large part around the entries, their structure means that individual portions of the documents—notably the chronology structures and Divs—can be indexed and retrieved to produce the dynamism that users seem to desire. We decided quite early that these would be irreducible units for search and delivery, although we are experimenting with permitting the return of abbreviated or "short form" search results which return less than an entire Div.

Altogether, the Orlando Delivery System 1.0 represents the first fruits of our efforts to produce a new kind of user-oriented literary history. The tagging has indeed provided the basis for dynamically produced sets of informative and interpretive text about women's writing in the British Isles. The delivery system makes the structuring and retrieval principles evident to the user, without letting them overwhelm, so that the conclusions users take away from their engagement with the materials can be informed by an awareness of the tagging. Results produced by the ODS are in fact the outcome of a dynamic relationship between user interests and the tagging, the results of a new kind of intellectual inquiry. Orlando's tagging scheme is so interpretive that it will ideally shift the focus of user attention away from simplistic notions of information retrieval and towards consideration of the processes of organizing, representing, and interpreting materials in electronic form, with the way literary history is reconstituted in this new medium. We need urgently to engage scholars who are not computing humanists in the debate over how scholarly work is represented electronically. The public "face" of Orlando aims to help our intellectual community face the challenges of the move to digital forms of scholarly text.

Works Cited:

Applied Research in ‘Patacriticism. University of Virginia. http://www.iath.virginia.edu/~jjm2f/arp/home.html

McGann, Jerome. “Culture and Technology: The Way We Live Now, What Is To Be Done?” North American Victorian Studies Association New Frontiers Conference, University of Toronto, October 2004. Forthcoming in New Literary History.

The Orlando Project. http://www.ualberta.ca/ORLANDO/

The Text Encoding Initiative. http://www.tei-c.org/

Disseminating Research in Accessible Ways: the Globalization and Autonomy Compendium

William Coleman, McMaster University
Andrew Mactavish, McMaster University

The *Globalization and Autonomy* Research project is funded under SSHRCC's Major Collaborative Research Initiatives program. The research team is looking at the dialectical relationships and interplay between globalization and autonomy. The research group is a large one involving about 40 co-investigators in 12 universities across Canada, another 20 or so academic contributors from outside Canada, a growing group of doctoral and postdoctoral scholars, and an affiliated interdisciplinary research team of 12 scholars based in Tunisia, but including some members from Spain and France as well.

In addition to preparing a number of academic volumes, the project team made a commitment to SSHRCC to make its findings available to a broad public audience. Working with Geoffrey Rockwell and Andrew Mactavish as technical editors, the team has worked toward building an on-line compendium of its findings. Using the most advanced standards for text editing on the web, the Compendium would include summaries of 750 to 1000 words of each of the chapters in the academic volumes. These summaries would be linked electronically to a glossary of key concepts and terms, events, places, organizations and individual persons. Finally, both of these parts would be linked in turn to a comprehensive bibliography of sources for the research.

LetSum, an Automatic Text Summarization system in Law field

Atefeh Farzindar and Guy Lapalme, RALI, Département d'Informatique et recherche opérationnelle, Université de Montréal

This paper presents our work on the development of a new methodology for automatic summarization of justice decision. We describe LetSum (Legal text Summarizer), a prototype system, which determines the thematic structure of a judgment in four themes Introduction, Context, Juridical Analysis and Conclusion. Then it identifies the relevant sentences for each theme. We have carried our an evaluation of produced summaries with statistical method and also human evaluation based on jurist judgment. The results so far indicate good performance of the system when compared with other summarization technologies.

1 Introduction

Legal experts perform difficult and responsible legal clerical work which requires accuracy and speed. This task includes understanding, interpreting, explaining and researching in a wide variety of legal documents. A summary of a judgement, as a compressed but accurate restatement of its content, helps in organizing a large volume of documents and in finding the relevant judgments for their case. For this reason, the decisions are frequently manually summarized by legal experts. But the human time and expertise required to provide manual summaries for legal resources, makes human summaries very expensive. This problem shows the interest in automated text summarization to process the ever increasing amount of documents. In this research, within the framework of the project TAPoR at Université de Montréal, in collaboration with the law faculty and department of computer science, we have developed a summarization system, called LetSum, for producing short summaries for legal decision of the proceeding of a court. We have attempted to reproduce the results of human expert reasoning by organizing and manipulating factual and heuristic knowledge.

We present our approach for summarizing the legal record of the proceedings of federal courts in Canada and presenting it as a table-style summary for the needs of lawyers and experts in the legal domain. The FLEXICON (Smith and Deedman, 1987), SALOMON (Moens et al., 1999) and SUM (Grover et al., 2003) projects and (Borges et al., 2003) attest the importance of the exploration of legal knowledge for sentence categorisation and summarisation. Our method investigates the extraction of the most important units based on the identification of the thematic structure in the document and the determination of argumentative themes of the textual units in the judgment (Farzindar et al., 2004).

In Canada, the Canadian Legal Information Institute (CANLII) gathers legislative and judicial texts in order to make a virtual library of Canadian law accessible for free on the Internet (www.canlii.org). The large volume of legal information in electronic form creates a need for the creation and production of powerful computational tools in order to extract relevant information in a condensed form.

The lawyers need to process previous legal decisions to find a solution to a legal problem not directly indicated in the law, they look for *precedents* of similar cases. Each decision contains the reasons which justify the solution for a legal problem. They constitute a *law jurisprudence precedent* from which it is possible to extract a legal rule that can be applied to similar cases.

One reason for the difficulty of the work in legal field is the complexity of the domain: specific terminology of the legal domain and legal interpretations of expressions produce many ambiguities. For example, the word *disposition* means nature, effort, mental attitude or property in general English but in legal terms it means the final part of a judgment indicating the nature of a decision: acceptance of an inquiry or dismissal. That is why we have constructed our conceptual dictionary containing 200 concepts of legal domain. In this project we collaborate with legal experts of CanLII.

Thematic segments give information which can be used to answer specific questions about the theme of the segment such as: what are the themes of a document? How is a theme used? How are the problem and the facts presented? How does a judge reason to reach a conclusion? In this paper, we will describe how we deal with the problem of the exploration of structure of document and content selection, according to the themes of a judgement.

1.1 Legal text summarization

Our approach to produce the summary is based on the identification of the thematic structure to find the argumentative themes of the judgment. This approach is a result of our corpus analysis in which we compared model summaries written by humans with the texts of the original judgments. The textual units considered as important by the professional abstractors were aligned manually with one or more elements of the source text. We look for a match between the information considered important in the professional abstract and the information in the source documents.

We extract the relevant sentences for each theme and present them as a table-style summary. Showing the information considered important could help the user read and navigate easily between the summary and the source judgment. If a sentence seems more important for a user and more information is needed about this topic, the complete thematic segment containing the selected sentence could be presented.

The identification of these structures separates the key ideas from the details of a judgment and improves readability and coherency in the summary. Therefore, in the presentation of a final summary, we propose to preserve this organization of the structure of the text in order to build a table-style summary.

Our corpus contains 3500 judgments of the Federal Court of Canada, which are available in HTML on www.canlii.org/ca/cas/fct. For some of these decisions, their summaries written by professional legal abstractors are available. We analyzed manually 50 judgments in English as well as their human written summaries. These judgments were suggested by the lawyers of CanLII project, as representing the *standard* judgments with *ideal* summaries.

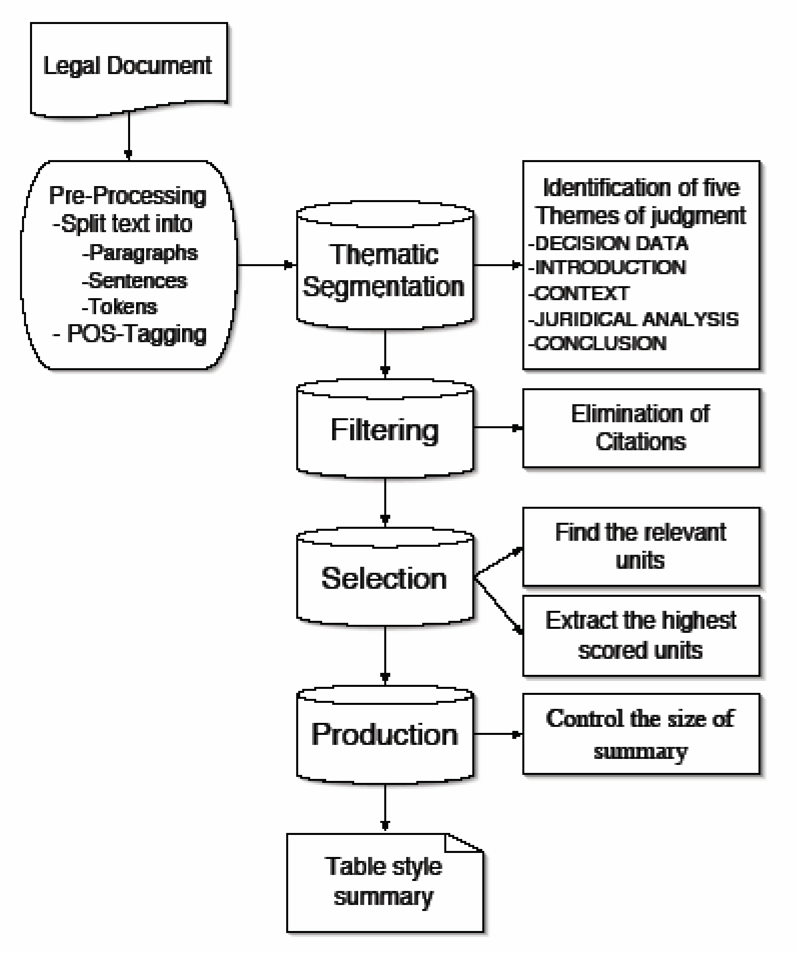


Figure 1: The procedural steps for generating of table-style summary

The rest of the corpus is used for statistical computations.

2 Components of LetSum

To process an input decision to LetSum, the system first does some pre-processing. The summary is built in four phases (Figure 1): thematic segmentation, filtering of less important units such as citations of law articles, selection of relevant textual units and production of the summary within the size limit of the abstract(Farzindar and Lapalme, 2004).

Pre-Processing splits the input judgment into main units. First the body of the text of the decision are identified. Some keywords like *Reasons for order, Reasons for judgment* and *order* separate the basic data (date, name of court, etc.), placed in the head of document, from the beginning of the judgment. The features used for the end of the decision are the date and place of hearing, name and signature of the judge. Then the document is divided into: section titles, paragraphs, sentences and tokens. To determine the Part-of-Speech tags, the tagger described by (Hepple, 2000) is used.

Thematic segmentation is based on the specific knowledge of the legal field. According to our analysis, the texts of jurisprudence have a thematic structure, independently of the category of judgment. Textual units dealing with the same subject form a thematic segment set. In this context, we distinguish four themes which divide the legal decisions into thematic segments, based on the experimental work of judge Mailhot (Mailhot, 1998):

Introduction describes the situation before the court and answers these questions: who? did what? to whom?

Context explains the facts in chronological order, or by description. It recomposes the story from the facts and events between the parties and findings of credibility on the disputed facts.

Juridical Analysis describes the comments of the judge and finding of facts, and the application of the law to the facts as found. For the legal expert this section of judgment is the most important part because it gives a solution to the problem of the parties and leads the judgment to a conclusion.

Conclusion expresses the *disposition* which is the final part of a decision containing the information about what is decided by the court.

For thematic segmentation the following information are used: the presence of significant section titles, the positions of a segment, the identification of direct or narrative style (as the transition of **Context** and **Juridical Analysis** segments), certain linguistic markers. We present the heuristics and some examples of linguistic markers for each theme:

Introduction presents a short description of a case with cue phrases such as: *application for judicial review, application to review a decision, motion filed by, Statement of Claim*. This segment is at the beginning of the judgement after a title like: *Reasons for order, Reasons for judgment* and *order*.
Significant section titles are *Introduction* and *Summary.*

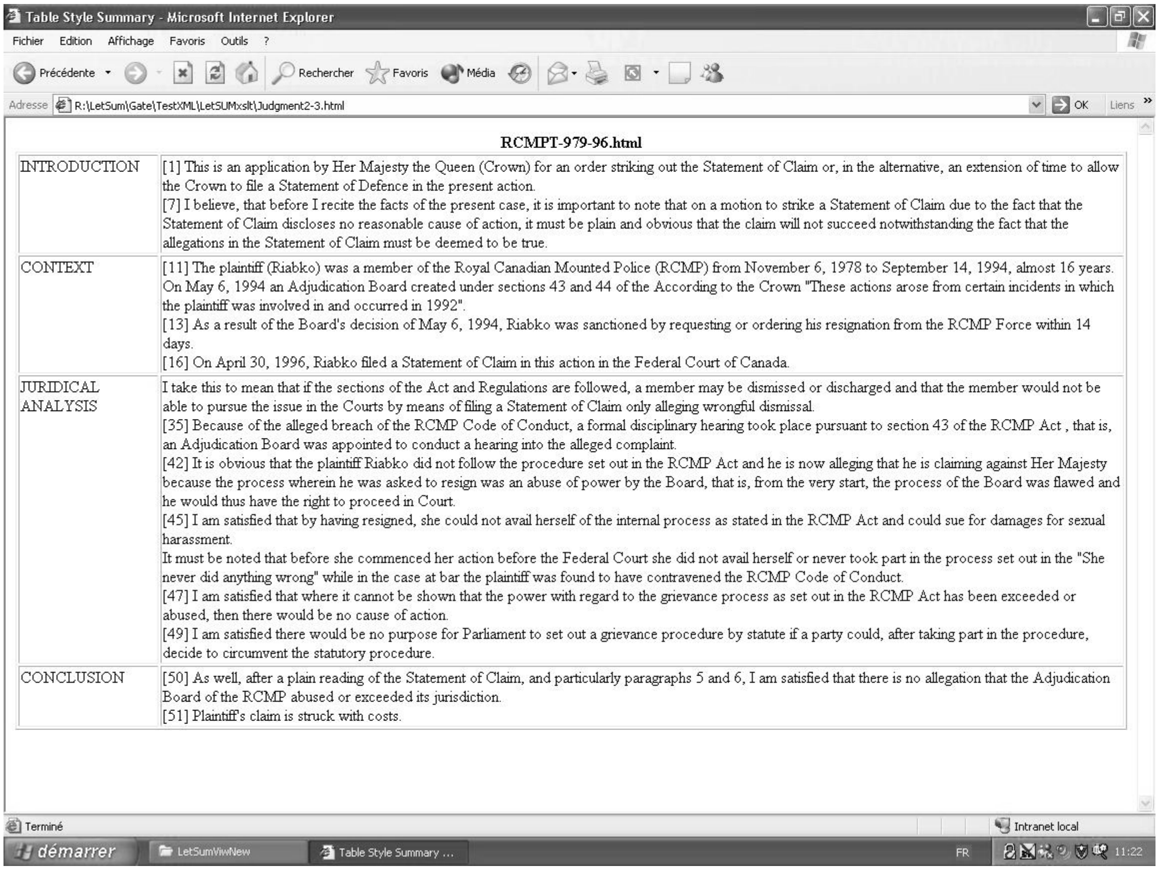
Context introduces the parties (eg. *The Applicant is a 52 year old citizen of Sri Lanka*). It describes the application request using markers such as: *advise, indicate, concern, request*, etc. (eg. *Motion concerns Air Canada's failure to provide ground services in the French language at the Halifax airport*). It explains the situation in the past tense and narration form. Section titles are *Facts, Background, Factual background* and *Agreed statement of the facts.*

Juridical Analysis, in which the judge gives his explanation on the subject thus the style of expression is direct using *I.*
Cue phrases are *In reviewing the sections No. of the Act, Pursuant to section No., As I have stated, In the present case, The case at bar is.* Section titles are *Analysis* and *Decision of the court.*

Conclusion contains the final result of the court decision using phrases such as: *The motion is dismissed, the application must be granted*. This segment is at the end of the judgment before the signature of the judge (judge's name, date, place of hearing, etc.). Section titles are *Conclusion, Costs* and *Disposition.*

Filtering identifies parts of the text which can be eliminated, without losing relevant information for the summary. In a judgment, citation units (sentence or paragraph) occupy a large volume in the text, up to 30%, of the judgment, whereas their content is less important for the summary (according to our annual alignments between summaries and sources). This is why we remove citations inside blocks of thematic segments. We thus filter two categories of segments: submissions and arguments that report the points of view of the parties in the litigation and citations related for previous issues or references to applicable legislation. In the case of eliminating a citation of a legislation (eg. law articles), we save the reference of the citation in **decision data** in the field of authority and doctrine.

The identification of citations is based on two types of markers: direct and indirect. A direct marker is one of the linguistic indicators that we classified into three classes: verbs, concepts (noun, adverb, adjective) and complementary indications. Examples of verbs of citation are: *conclude, define, indicate, provide, read, reference, refer, say, state, summarize.* Examples of the concepts are: *following, section, subsection,page, paragraph, pursuant.* Complementary indications



RCMPT-979-96.html

INTRODUCTION	[1] This is an application by Her Majesty the Queen (Crown) for an order striking out the Statement of Claim or, in the alternative, an extension of time to allow the Crown to file a Statement of Defence in the present action. [7] I believe, that before I recite the facts of the present case, it is important to note that on a motion to strike a Statement of Claim due to the fact that the Statement of Claim discloses no reasonable cause of action, it must be plain and obvious that the claim will not succeed notwithstanding the fact that the allegations in the Statement of Claim must be deemed to be true.
CONTEXT	[11] The plaintiff (Riabko) was a member of the Royal Canadian Mounted Police (RCMP) from November 6, 1978 to September 14, 1994, almost 16 years. On May 6, 1994 an Adjudication Board created under sections 43 and 44 of the According to the Crown "These actions arose from certain incidents in which the plaintiff was involved in and occurred in 1992". [13] As a result of the Board's decision of May 6, 1994, Riabko was sanctioned by requesting or ordering his resignation from the RCMP Force within 14 days. [16] On April 30, 1996, Riabko filed a Statement of Claim in this action in the Federal Court of Canada.
JURIDICAL ANALYSIS	I take this to mean that if the sections of the Act and Regulations are followed, a member may be dismissed or discharged and that the member would not be able to pursue the issue in the Courts by means of filing a Statement of Claim only alleging wrongful dismissal. [35] Because of the alleged breach of the RCMP Code of Conduct, a formal disciplinary hearing took place pursuant to section 43 of the RCMP Act , that is, an Adjudication Board was appointed to conduct a hearing into the alleged complaint. [42] It is obvious that the plaintiff Riabko did not follow the procedure set out in the RCMP Act and he is now alleging that he is claiming against Her Majesty because the process wherein he was asked to resign was an abuse of power by the Board, that is, from the very start, the process of the Board was flawed and he would thus have the right to proceed in Court. [45] I am satisfied that by having resigned, she could not avail herself of the internal process as stated in the RCMP Act and could sue for damages for sexual harassment. It must be noted that before she commenced her action before the Federal Court she did not avail herself or never took part in the process set out in the "She never did anything wrong" while in the case at bar the plaintiff was found to have contravened the RCMP Code of Conduct. [47] I am satisfied that where it cannot be shown that the power with regard to the grievance process as set out in the RCMP Act has been exceeded or abused, then there would be no cause of action. [49] I am satisfied there would be no purpose for Parliament to set out a grievance procedure by statute if a party could, after taking part in the procedure, decide to circumvent the statutory procedure.
CONCLUSION	[50] As well, after a plain reading of the Statement of Claim, and particularly paragraphs 5 and 6, I am satisfied that there is no allegation that the Adjudication Board of the RCMP abused or exceeded its jurisdiction. [51] Plaintiff's claim is struck with costs.

Figure 2: A table-style summary produced by LetSum, the original judgment has 3500 words and the summary is 15% of the source

include numbers, certain prepositions, relative clauses and typographic marks (colon, quotation marks). The indirect citations are the neighboring units of a quoted phrase. For example, citation segment in the phrase *paragraph 78(1), which reads as follows:* is identified using direct markers but it points to the textual units with no direct marker which are also quotations. We thus identify the enumerated sentences following a quoted sentence for determining a group of citations.

Selection builds a list of the best candidate units for each structural level of the summary. LetSum computes a score for each sentence in the judgment based on heuristic functions related to the following information: position of the paragraphs in the document, position of the paragraphs in the thematic segment, position of the sentences in the paragraph, distribution of the words in document and corpus ($tf \cdot idf$).Depending on the given information in each layered segment, we have identified some cue words and linguistic markers. The thematic segment can change the value of linguistic indicators. For example, the phrase *application is dismissed* that can be considered as an important feature in the **conclusion** might not have the same value in **context** segment. At the end of this stage, the passages with the highest resulting scores are sorted to determine the most relevant ones.

Production of the final summary controls the size of the summary and displays the selected sentences in tabular format. The final summary is about 10% of the source document. The elimination of the unimportant sentences takes into account length statistics based on our observation from human abstracts.

System ID	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
LetSum	**0.57500**	**0.31381**	**0.20708**	**0.15036**	**0.45185**
Baseline	0.47244	0.27569	0.19391	0.14472	0.34683
Mead	0.45581	0.22314	0.14241	0.10064	0.32089
Word	0.44473	0.21295	0.13747	0.09727	0.29652
P. Mining	0.32833	0.15127	0.09798	0.07151	0.22375

Table 1: Result of statistical evaluation with ROUGE, LetSum is the first with the best evaluation scores

In the **Introduction** segment, units with the highest score are kept within 10% of the size of the summary. In the **Context** segment, the selected units occupy 25% of the summary length. The contribution of the **Juridical Analysis** segment is 60% and the units with the role **Conclusion** occupy 5% of the summary.

Figure 2 shows an example of a table-style summary generated by LetSum. The summary is 15% of the source judgment.

3 Evaluating LetSum

We evaluate LetSum in two steps; first the evaluation of the modules of the system and second, a global evaluation of produced summaries. The evaluations of components of LetSum are very promising; we obtained 90% correct segmentation for thematic segmentation module and 97% correct detection for filtering stage (correct detection of 57 quoted segments over 60).

The final global evaluation includes the intrinsic and extrinsic tests (Spark-Jones and Galliers, 1995). Intrinsic evaluations test the system in of itself and extrinsic evaluation test the system in relation to some other task.

For intrinsic evaluations, we conducted a pilot study on 10 long judgments (average 8 pages). We compare the summaries produced by machine with the reference summaries written by professional abstractors. This evaluation is Recall-based, which measures how many of the reference summary sentences the machine summary contains. For measuring recall, we used the ROUGE software (Lin, 2004), which determines the quality of a summary by comparing it to ideal summaries created by humans. The score of ROUGE-N is based on the number of n-grams occurring at the reference summary side. For example, ROUGE-2 computes the number of two successive words occurring between the machine summary and ideal summary. For measuring ROUGE-L, we view a summary sentence as a sequence of words. This evaluation computes the longest common subsequence of words to estimate the similarity between two summaries. We compared LetSum with some other systems. A baseline system is a simple system which other system can be compared. Traditionally, for a newspaper article a baseline is the first paragraphs of the text. In our case, we defined a baseline with compression rate of 15% of the source document with the following algorithm:

- Choose 8% words of the beginning of the judgment. According our thematic segmentation, it takes the sentences from the themes **Introduction** and **Context**. If the last sentence is cut with this limit, complete it.
- Choose the last 4% words of the judgment with themes **Juridical Analysis** and **Conclusion**. If the first sentence is cut with this limit, complete it.

For automatic evaluation, we have compared the produced sentences by LetSum with: the baseline, the commercial automatic summaries produced by Microsoft Word and Pertinence Mining (www.pertinence.net) and a state-of the-art summarization system Mead (Radev et al., 2003), with the human reference summaries. Table 1 shows the result of this evaluation. The higher score means better score and more performance system. LetSum is ranked first with the best evaluation scores.

The evaluation results show the interest of developing a summarization system for a specific domain because it is more and more difficult to generate a general summary without consideration of the user profile and the domain.

Our extrinsic evaluations will be based on legal expert judgment. We have defined a series of specific questions for the judgment (with the help a lawyer of CanLII), which cover the main topics of the document. If a user is able to answer the questions correctly by only reading the summary, it means the summary contains all the necessary information of the source judgment. We are currently performing our human based evaluations.

4 Conclusion

LetSum is a one the few systems developed specifically for summarization of legal documents. This system is implemented in an environment such as CanLII which has to deal with thousand of texts and produce summaries for each. We have presented our approach based on the extraction of relevant units in the source judgment by identifying the document's structure and determining the themes of the segments in the decision. The generation of the summary is done in four steps: thematic segmentation to detect the legal document structure in four themes **Introduction, Context, Juridical analysis** and **Conclusion**, filtering to eliminate unimportant quotations and noises, selection of the candidate units and production of table-style summary. The presentation of the summary is in a tabular form along the themes of the judgment.

The evaluation of the system includes the intrinsic and extrinsic tests. The result of intrinsic evaluation of LetSum is very promising. We are completing the extrinsic evaluation based on legal expert judgments.

Acknowledgements

We would like to thanks LexUM group of legal information-processing laboratory of the Public Law Research Center at the University of Montréal for their valuable suggestions. This work has been developed using infrastructures funded by TAPoR. At Université de Montréal, TAPoR is funded by the Canadian Foundation for Innovation Project 5637 and the Qubec Ministry of Education. This project supported by Natural Sciences and Engineering Research Council of Canada.

References

Filipe Borges, , Raoul Borges, and Danièle Bourcier. 2003. Artificial neural networks and legal categorization. In *The 16th Annual Conference on Legal Knowledge and Information Systems (JURIX'03)*, page 187, The Netherlands, 11 and 12 December.

Atefeh Farzindar and Guy Lapalme. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out Workshop held in conjunction with ACL'2004*, pages 27–34, Barcelona, Spain, 25–26 July.

Atefeh Farzindar, Guy Lapalme, and Jean-Pierre Desclés. 2004. Résumé de texts juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte : solutions et perspectives*, 45(1):26 pages.

Claire Grover, Ben Hachey, and Chris Korycinski. 2003. Summarising legal texts: Sentential tense and argumentative roles. In Dragomir Radev and Simone Teufel, editors, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03),* pages 33–40, Edmonton, Alberta, Canada, May 31 - June 1.

Mark Hepple. 2000. Independence and commitment: Assumptions for rapid training and execution of rule-based part-of-speech taggers. In *the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 278–285, October.

Chin Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out Workshop held in conjunction with ACL'2004,* pages 74–81, Barcelona, Spain, 25–26 July.

Louise Mailhot. 1998. *Decisions, Decisions: a handbook for judicial writing*. Editions Yvon Blais, Québec, Canada.

Marie-Francine Moens, C. Uyttendaele, and J. Dumortier. 1999. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2):151–161.

Dragomir Radev, Jahna Otterbacher, Hong Qi, and Daniel Tam. 2003. Mead reducs: Michigan at duc 2003. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1. Association for Computational Linguistics.

J. C. Smith and Cal Deedman. 1987. The application of expert systems technology to case based law. *ICAIL*, pages 84–93.

Karen Spark-Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

Text analysis and the problem of pedantry

Julia Flanders, Women Writers Project, Brown University

Computer-aided text analysis has a long history as one of the earliest applications of information technology to humanities research. However, although the rise of the web has brought digital resources to the fore, text analysis has not been widely adopted as a technique of literary study in the academic mainstream. We can find some explanation for current academic resistance to these methods by examining historical debates about the role of the scholar (and his devalued doppelganger, the pedant) and the kinds of literary evidence and knowledge that should inform scholarly work. The tools and research methods now being used for computer-aided literary analysis must work against the grain of this resistance and find a new way of articulating the role of the detail and the datum within literary interpretation.

Automated text categorization: theory and application to computer-assisted text analysis in the humanities

Dominic Forrest, Laboratoire d'Analyse Cognitive de l'Information, Université du Québec à Montréal

The use of computer-assisted text analysis techniques in the humanities has become an important research field, especially since the last decade (Hockey, 2000; Popping, 2000). Many most promising applications of computer-assisted text analysis in the humanities have been presented and are still the object of important research programs. Most of these applications concern the retrieval of textual information, the automatic indexing of corpus, the automatic summarization of corpora, the automatic hypertext generation within documents, etc. Recently, many researchers have concentrated their efforts on the application of categorization techniques to document analysis (Jackson and Moulinier, 2002; Sebastiani, 2002). Our presentation will be divided into two parts. In the first one, we will present some of the most important techniques used in automatic categorization. In the second part, we will present a categorization technique based on the automatic emergence of categories. Finally, we will see how automated categorization techniques can be applied successfully to computer-assisted thematic analysis of documents.

Computer Text Analysis and the Man from Wausau

Paul A. Fortier, Centre on Aging/Department of French, Spanish and Italian, University of Manitoba

At the 2004 ALLC/ACH Congress in Goteborg, Joseph Rudman noted with dismay the lack of influence that even first-rate computer text analysis, like the Mosteller and Wallace study of *The Federalist Papers*, have had on mainline humanities studies. Thomas Corns made a similar statement in 1990 at the Siegen ALLC/ACH Congress. It is possible to find many reasons for this phenomenon, including the rise of post-structuralist denial of objective reality at the juncture when computer text analysis was becoming possible, and the "allergic" reaction that many humanities scholars have to anything that smacks of technology or mathematics.

It must also be recognized that computer-oriented humanities scholars have all too often chosen to concentrate their efforts on what could most easily be processed by computer, producing results which make no sense in terms of the realities with which scholars are familiar. A classic example of this is stylometry, with its counting of type/token ratios, and the frequency of words appearing only once in a text. Not only is it difficult to find a coherent explanation of why these measures are significant. They are also all too frequently done is terms of spelled forms. Everything is done to facilitate computer processing, no messy time-wasting by lemmatising texts is contemplated, and most-impressive statistics are produced. But are the statistics impressive? Étienne Brunet's work on French literature is a case in point. His massive study of the language of Victor Hugo shows a table on vocabulary richness, based on type/token ratios, in which 21 out of 22 texts are significant at the .01 level (Brunet I, 26). As for the number of words appearing only once in their text, the results published by Brunet (I, 36) show that these numbers are significant at the .01 level for 20 out of 22 texts. One does not have to be anti-numerate to find such results perverse. This is a clear case of computing facility taking precedence over understandable analysis.

For an example of literary analysis, take it that one has read André Gide's novel *L'Immoraliste* and on the basis of that reading, one feels that twenty-one themes are important for understanding the text. According to current theories of semantics (cf. Lakoff), anyone who knows the language would be able to identify the words related to the themes, and a computer-generated concordance or word list would certainly facilitate this effort. Further, it would not be a large effort to look up these words in a frequency dictionary like Engwall's *Vocabulaire du roman français*. From there it is simple to see if these themes are used more frequently in Gide's novel than in Engwall's representative sample of the genre.

In point of fact, the language contains, in all probability, more words evoking a theme like *amour* or *santé* than those found in Gide's text. If one wants to measure the extent to which Gide's word usage is extraordinarily high for a given theme, one really should do it in terms of all the words which are potentially related to the theme and which Gide might have used, rather than just the ones he did use. Table 1 shows that this reaction is correct, in each and every case the words used by Gide (found in column four) are fewer in number than the words suggested by the synonym dictionaries (see column two). In fact, the number of different words evoking each of the themes is larger in Engwall's sample of twenty-five different novels, than in Gide's novel, but smaller than what was suggested by the synonym dictionaries. This confirms the widely-held view that literary language is a subset of the language in general, and that what is found in an individual literary text is a subset of literary language.

As far as determining whether the themes identified by close reading are really important according to an objective measure, Z-scores in terms of a Poisson distribution are a reasonable test (see Table 2). The Z-scores are derived in the following fashion. The frequency for all the words evoking the theme shown in Engwall (col. 2) is divided by 500 000 (the total number of words sampled) and multiplied by 41,178 (the number of words in *L'Immoraliste*) to arrive at a predicted frequency (***pred.***). If the vocabulary of Gide's text is what could be expected from Engwall's data, the observed frequency of a theme will be close to this predicted value as in the case of *claustration* and *eau*. Within the chosen model, the standard deviation is defined as the square root of the predicted value, and is shown in column six (***s.d.***) of Table 2. The Z-score is the difference between the observed and the predicted values (***dev.***), divided by the standard deviation (***s.d.***).

Usually a Z-score of two or greater is considered statistically significant. The imperfect independence of the categories, as well as the fact that the techniques used do not record examples of metaphor, irony and other common literary devices, justfiy considering a Z-score of three, corresponding to a probability of approximately 1 in 800 in a one-tailed test, to be the break-point for statistical significance in this material (Fortier 293-94).

Even with this more stringent requirement, 15 of the 21 themes studied are more frequent than predicted to a statistically significant extent. The applicable Z-scores are in bold in Table 2. The probability of such a high number of statistically significant observations occurring by chance alone is virtually nil, if the statistical model is appropriate. It is thus demonstrated that the choice of these themes as important leads to a valid description of the underlying literary reality, one which cannot be ascribed to the random action of the critic's imagination.

As a matter of interest, the extreme right-hand column of Table 2 also shows Z-scores based on predicted values of just the words used by Gide related to each theme (these predicted values and their standard deviations are not shown in the table), not on the predicted values of the whole groups of words suggested by the synonym dictionaries. These results are preceded by a *, following the convention used in linguistics for examples known to be erroneous. In all cases, the Z-scores are considerably higher than those computed on the basis of the full vocabulary related to the theme. In fact, the mean increase is greater than three. These results confirm the necessity to evaluate frequencies for the whole set of words potentially evoking a given theme, as has been done, rather than just the words appearing in a single text if one wants to avoid the danger of falsely positive results. Here again doing the extra work required to produce results justifiable in terms of the data being studied is preferable to a less onerous but more superficial approach.

My final example was reported by Peyawary. English comes in three main dialects: American, British, Indian. Each of these dialects has a million-word corpus of various genres designed to be representative of that dialect of English as a whole. Teaching English as a second language is usually done in terms of a core vocabulary of two thousand words determined in the early twentieth century to be the essential of the language. There are several reasons for wanting to redo this designation of core vocabulary, not the least of which is the bias towards personal hygiene and Christianity found in the original list.

Since the frequency dictionaries exist, it might seem to be a simple matter to determine what is the core vocabulary by sorting the frequency lists into descending order of frequency and comparing them using Spearman's rank correlation coefficient. When the frequencies are no longer statistically similar at the .05, level it will be possible to say that regional divergences have taken precedence over the fundamental language. On the basis of this method, it is shown that the core vocabulary of English consists of 131 different words! Such results are perverse, because they demonstrate that a speaker of British English could not understand a software manual written in New York or a novel published in Calcutta. We know that is not the case.

We also know that any language, even after one has reduced morphological variation by converting all words to their dictionary forms, consists of many sub-systems of which the syntactic and the semantic are the most familiar. When one takes into account the syntactic system by coding words with their part of speech, a quite considerable task which involves examining, in this case, the context of three million words, one can compare the frequencies of prepositions in each dialect, then the frequencies of auxiliary verbs, then of adverbs, and so on. This approach, which integrates the syntactic structure of language, produces a core vocabulary of 1061 words, covering 74.1% of the vocabulary in the three frequency corpora studied, which will certainly improve the teaching of English as a second language.

In all of these cases it has been shown that with a little extra effort, guided by what is already known about the phenomena being studied, one can produce results that make sense in terms of the discipline in question, results that have at least a better chance of being taken seriously by mainstream scholars.

Acknowledgements

My research has been supported by many research and travel grants provided by the University of Manitoba, by many grants from the Social Sciences and Humanities Research Council of Canada, of which the most recent one is 410-2003-500, and by the Alzheimer Society of Canada, grant number 04-36.

Theme	Synonym Dictionaries	Engwall	Gide, *Immoraliste*
Amour	245	137	72
Beauté	237	129	61
Bien	195	133	67
Chaleur	260	107	33
Claustration	54	36	11
Eau	410	179	69
Ennui	211	132	68
Faiblesse	234	116	57
Force	367	219	94
Froid	105	41	17
Joie	200	119	55
Laideur	117	69	32
Lumière	179	83	27
Maladie	1524	282	79
Mort	164	76	20
Nuit	129	54	23
Peur	114	65	36
Santé	75	36	22
Végétation	1033	285	72
Vie	361	156	80
Violence	274	172	66

***Table 1:* Number of Words in Themes**

References

Brunet, Étienne. *Le Vocabulaire de Victor Hugo.* 3 vols. Paris: Champion, 1988.

Engwall, Gunnel. *Vocabulaire du roman français (1962-68): Dictionnaire des Fréquences.* Data Linguistica. Stockholm: Almqvist & Wiksell, 1984.

Fortier, Paul A. "Some statistics of themes in the French novel." *Computers and the Humanities* 23:293-99.

Gide, André. *L'Immoraliste.* 1902. *Romans, récits, soties, oeuvres lyriques.* Eds. Y. Davet & J.J. Thierry. Bibliothèque de la Pléiade. Paris: Gallimard, 1958.

Lakoff, George. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind.* Chicago, University of Chicago Press, 1987.

Peyawary, Ahmad S. *The Core Vocabulary of International English: A Corpus Approach.* Bergen: HIT Centre. 1999.

The Collected Letters of Bertrand Russell: An Electronic Edition

Nicholas Griffin, McMaster University
James Chartrand, Open Sky Solutions

Russell's complete correspondence is very, very large; so large that it not only makes a print edition impossible, but has made it worthwhile to create an editing programme specially for the project. The programme, created by James Chartrand, covers the entire editiorial process from the initial capture of the image of a letter from an archival security microfilm, through the transcription and annotating of the letter, to its final approval and inclusion in the edition's on-line database. The programme applies TEI-conformant tags to the text automatically, eliminating errors, inconsistency, and the need for transcribers and editors to be extensively trained in the ways of the TEI. The programme makes possible a project which, otherwise, would forever be beyond the resources of the Bertrand Russell Research Centre. Although specially created for the edition of Russell's letters, the programme should be adaptable to other editing projects, especially those which involve very large numbers of documents.

(De)Facing the Text: Irradiated Textuality and Deformed Interpretations

David L. Hoover, English Department, New York University

My own recent work in statistical stylistics, corpus stylistics, and authorship attribution relies heavily on electronic texts and computational methods, and would be impossible without them. Here I concentrate on less computationally intensive interpretive effects of electronic texts. Specifically, I argue that McGann's Radiant Textuality (2001), shaped by both literary theory and the nature of electronic texts, is often counter-productive and wrong-headed, and that McGann's valorization of the instability of (electronic) texts and interpretation lead to poor results. Among many tempting examples in Radiant Textuality, I concentrate on two: the scanner experiment that leads to the dictum that "no text is self-identical," and the "deformance" of Stevens's "The Snowman."

McGann scans an advertizement page from a Victorian periodical and subjects it to OCR repeatedly. Not surprisingly, given the age and complexity of the document, this yields documents with different zones and some variation in the alphanumeric text (he doesn't say how much) (144-46). It also leads him to claim that all texts are marked/interpreted, and that texts are not meaning containers, but rules for self-generation. These valuable insights are followed by the more radical claim that the scanner experiment constitutes "a powerful physical argument" that "no text is self-identical" (145). This skeptical argument is finally unanswerable except by kicking a text down a hill, but textual instability must be kept in perspective. McGann originally planned similar operations on "a relatively straightforward piece of prose formatted margin-to-margin in standard block form" (144). My own experiments with such texts suggest that the non-self-identity of McGann's texts is less profound than he suggests. Paper texts and their scanned electronic counterparts are remarkably stable, and their remaining instability is predictable and not very important.

More central to McGann's project is the "deformation" of poems, especially Wallace Stevens's "The Snow Man." The argument is that deformance is liberating and revealing, and that it "clarifies the secondary status of the interpretation" (120). (An electronic text of the poem makes deformance easier, but the computer is not an important part of the process.) If he had read less theory and more stylistics, McGann might have noted the long rich tradition of text- alteration, most notably exemplified by Rob Pope's wonderful Textual Intervention (1995), but his deformances lead more often to nonsense than insight. For example, he claims that printing only the nouns in their original positions "enhances the significance of the page's white space, which now appears as a poetic equivalent for the physical "nothing" of snow" (123). But the white space doesn't belong to the poem, only to the deformed version, and similar white spaces occur in any poem printed like this, whether or not it mentions whiteness, nothingness, or snow. Perhaps printing only the nouns from "The Raven" would allow the black type of the words to appear as the poetic equivalent of the night's plutonian shore? That way lies madness, not illumination.

One of the most egregiously wrong-headed claims McGann makes brings us back to electronic texts and to computational methods. Once the poem's nouns are printed by themselves, he argues that this deformance shows the poem to be noun-heavy and noun-balanced, with two nouns typically appearing in each line. The counting of the nouns itself is questionable, but the biggest problem is that, without a norm for comparison, any claim that the poem is noun-heavy is meaningless. The existence of large electronic corpora of poetry has finally made possible the determination of reasonably persuasive norms, but even a much more modest hand-count shows that "The Snow Man" is not noun-heavy, but noun-average.

Here and elsewhere, more attention to the text, more trust in its intersubjective stability, less concentration on the remaining areas of instability, more use of computational methods, and less celebration of the "freedom" of interpretation can yield more insights into literature and into the nature and practice of interpretation.

Proving and Improving Authorship Attribution Technologies

Patrick Juola and **John Sofko,** Duquesne University

Who wrote Primary Colors? Can a computer help us make that call? Despite a century of research, statistical and computational methods for authorship attribution are neither reliable, well regarded, widely used, or well understood. This paper presents a survey of the current state of the art as well as a framework for uniform and unified development of a tool to apply the state of the art, despite the wide variety of methods and techniques used.

1 Introduction

Determining the author of a particular piece of text has been a methodological issue for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers. In recent years, the development of improved statistical techniques (Holmes, 1994) in conjunction with the wider availability of computeraccessible corpora (Nerbonne, 2004) has made the automatic inference of authorship (variously called "authorship attribution" or more generally "stylometry") at least a theoretical possibility, and research in this area has expanded tremendously. From a practical standpoint, acceptance of this technology is dogged by many issues — epistemological, technological, and political — that limit and in some cases prevent its wide acceptance.

This paper presents a framework for development and analysis to address these issues. In particular, we discuss two major usability concerns, accuracy and userfriendliness. In broad terms, these concerns can only be addressed by expansion of the number of clients (users) for authorship attribution technology. We then present a theoretical framework for description of authorship attribution to make it easier and more practical for the development and improvement of genuine offtheshelf attribution solutions.

2 Background

With a history stretching to 1887 (Mendenhall, 1887), and 3,520 hits on Google[1], it is apparent that statistical/quantitative authorship attribution is an active and vibrant research area. With nearly 120 years of research, it is surprising that it has not been accepted by relevant scholars : "Stylometrics is a field whose results await acceptance by the world of literary study by and large."[2] This can be attributed at

[1] Phrasal search for "authorship attribution," July 30, 2004

[2] Anonymous, personal communication to Patrick Juola, 2004

least partially to a limited view of the range of applicability, to a history of inaccuracy, and to the mathematical complexity (and corresponding difficulty of use) of the techniques deployed.

For example, and taking a broad view of “stylometry” to include the inference of group characteristics of a speaker, the story from Judges 12:5–6 describes how tribal identity can be inferred from the pronunciation of a specific word (to be elicited). Specifically, the Ephramites did not have the /sh/ sound in their dialect, and thus pronounced words with such sounds differently than the Gileadites. A more modern version of such *shibboleths* could involve specific lexical or phonological items; a person who write of sitting on a “Chesterfield” is presumptively Canadian, and an older Canadian at that (Easson, 2002). (Wellman, 1936)[p. 114] describes how an individual spelling error — an idiosyncratic spelling of “toutch” was elicited and used in court to validate a document for evidence.

At the same time, such tests cannot be relied upon. Idiosyncratic spelling or not, the word “touch” is rather rare [86 tokens in the millionword Brown corpus (Kuˇcera and Francis, 1967)], and although one may be able to elicit it in a writing produced on demand, it’s less likely that one will be able to find it independently in two different samples.

People are also not consistent in their language, and may (mis)spell words differently at different times; often the tests must be able to handle distributions instead of mere presence/absence judgements. Most worryingly, the tests themselves may be inaccurate [see especially the discussion of CUSUM (Farringdon, 1996) in (Holmes, 1998)], rendering any technical judgement questionable, especially if the test involves subtle statistical properties such as “vocabulary size” or “distribution of function words,” concepts that may not be immediately transparent to the lay mind.

Questions of accuracy are of particular importance in wider applications such as law. The relevance of a document (say, an anonymously libelous letter) to a court may depend not only upon who wrote it, but upon whether or not that authorship can be demonstrated. Absent eyewitnesses or confessions, only experts, defined by specialized knowledge, training, experience, or education, can offer “opinions” about the quality and interpretation of evidence. U.S. law, in particular, greatly restricts the admissibility of scientific evidence via a series of epistemological tests[3] . The *Frye* test states that scientific evidence is admissible only if “generally accepted” by the relevant scholarly community, explicitly defining science as a consensus endeavor. Under *Frye*, (widespread) ignorance of or unfamiliarity with the techniques of authorship attribution would be sufficient by itself to prevent use in court. The *Daubert* test is slightly more epistemologically sophisticated, and establishes several more objective tests, including but not

[3] Frye vs. United States, 1923; Daubert vs. Merrill Dow, 1993.

limited to empirical validation of the science and techniques used, the existence of an established body of practices, known standards of accuracy (including socalled type I and type II error rates), a pattern of use in nonjudicial contexts, and a history of peer review and publication describing the underlying science.

At present, authorship attribution cannot meet these criteria. Aside from the question of general acceptance (the quote presented in the first paragraph of this section, by itself, shows that stylometrics couldn't pass the *Frye* test), the lack of standard practices and known error rates eliminates stylometry from *Daubert* consideration as well.

3 Recent developments

To meet these challenges, we present some new methodological and practical developments in the field of authorship attribution. In June 2004, ALLC/ACH hosted an "Adhoc Authorship Attribution Competition"(Juola, 2004a) as a partial response to these concerns. Specifically, by providing a standardized test corpus for authorship attribution, not only could the mere ability of statistical methods to determine authors be demonstrated, but methods could further be distinguished between the merely "successful" and "very successful." (From a forensic standpoint, this would validate the science while simultaneously, establishing the standards of practice and creating information about error rates.) Contest materials included thirteen problems, in a variety of lengths, styles, genres, and languages, mostly gathered from the Web but including some materials specifically gathered to this purpose. Two dozen research groups participated by downloading the (anonymized) materials and returning their attributions to be graded and evaluated against the known correct answers.

The specific problems presented included the following:

- *Problem A* (English) Fixedtopic essays written by thirteen Duquesne students during fall 2003.
- *Problem B* (English) Freetopic essays written by thirteen Duquesne students during fall 2003.
- *Problem C* (English) Novels by 19th century American authors (Cooper, Crane, Hawthorne, Irving, Twain, and 'noneoftheabove'), truncated to 100,000 characters.
- *Problem D* (English) First act of plays by Elizabethan/Jacobean playwrights (Johnson, Marlowe, Shakespeare, and 'none-of-the-above').
- *Problem E* (English) Plays in their entirety by Elizabethan/Jacobean playrights (Johnson, Marlowe, Shakespeare, and 'noneoftheabove').
- *Problem F* ([Middle] English) Letters, specifically extracts from the Paston letters (by Margaret Paston, John Paston II, and John Paston III, and 'noneoftheabove' [Agnes Paston]).
- *Problem G* (English) Novels, by Edgar Rice Burrows, divided into "early" (pre1914) novels, and "late" (post1920).
- *Problem H* (English) Transcripts of unrestricted speech gathered during committee meetings, taken from the *Corpus of Spoken Professional American-English*.

- *Problem I* (French) Novels by Hugo and Dumas (pere).
- *Problem J* (French) Training set identical to previous problem. Testing set is one *play* by each, thus testing ability to deal with crossgenre data.
- *Problem K* (SerbianSlavonic) Short excerpts from *The Lives of Kings and Archbishops*, attributed to Archbishop Danilo and two unnamed authors (A and B). Data was originally received from Alexsandar Kostic.
- *Problem L* (Latin) Elegaic poems from classical Latin authors (Catullus, Ovid, Propertius, and Tibullus).
- *Problem M* (Dutch) Fixedtopic essays written by Dutch college students, received from Hans van Halteren.

The contest (and results) were surprising at many levels; some researchers initially refused to participate given the admittedly difficult tasks included among the corpora. For example, Problem F consisted of a set of letters extracted from the Paston letters. Aside from the very real issue of applying methods designed/tested for the most part for modern English on documents in Middle English, the size of these documents (very few letters, today or in centuries past, exceed 1000 words) makes statistical inference difficult. Similarly, problem A was a realistic exercise in the analysis of student essays (gathered in a freshman writing class during the fall of 2003) – as is typical, no essay exceeded 1200 words. From a standpoint of literary analysis, this may be regarded as an unreasonably short sample, but from a standpoint both of a realistic test of forensic attribution, as well as a legitimately difficult problem for testing the sensitivity of techniques, these are legitimate.

Results from this competition were heartening. ("Unbelievable," in the words of one contest participant.) The highest scoring participant was the research group of Vlado Keselj, with an average success rate of approximately 69%. (Juola's solutions, in the interests of fairness, averaged 65% correct.) In particular, Keselj's methods achieved 85% accuracy on problem A and 90% accuracy on problem F, both acknowledged to be difficult and considered by many to be unsolvably so.

More generally, all participants scored significantly above chance. Perhaps as should be expected, performance on English problems tended to be higher than on other languages. Perhaps more surprisingly, the availability of large documents was not as important to accuracy as the availability of a large number of smaller documents, perhaps because they can give a more representative sample of the range of an author's writing. Finally, methods based on simple lexical statistics tended to perform substantially worse than methods based on Ngrams or similar measures of syntax in conjunction with lexical statistics. We continue to examine the detailed results in an effort to identify other characteristics of good solutions.

4 New technologies

The variation in these techniques can make authorship attribution appear to be an unorganized mess, but it has been claimed that under an appropriate theoretical framework (Juola, 2004b), many of these techniques can be unified, combined, and deployed. The initial observation is that, broadly speaking, all known human languages can be described as an unbounded sequence chosen from a finite space of possible events. For example, the IPA phonetic alphabet (Ladefoged, 1993) describes an inventory of approximately 100 different phonemes; a typewriter shows approximately 100 different Latin1 letters; a large dictionary will present an English vocabulary of 50–100,000 different words. An (English) utterance is "simply" a sequence of phonemes (or words).

The proposed framework postulates a threephase division of the authorship attribution task, each of which can be independently performed. These phases are :

- Canonicization — No two physical realizations of events will ever be exactly identical. We choose to treat similar realizations as identical to restrict the event space to a finite set.
- Determination of the event set — The input stream is partitioned into individual non-overlapping "events." At the same time, uninformative events can be eliminated from the event stream.
- Statistical inference — The remaining events can be subjected to a variety of inferential statistics, ranging from simple analysis of event distributions through complex patternbased analysis. The results of this inference determine the results (and confidence) in the final report.

As an example of how this procedure works, we consider a method for identifying the language in which a document is written. The statistical distribution of letters in English text is wellknown [see any decent cryptography handbook, including (Stinson, 2002)]. We first canonicize the document by identifying each letter (an italic *e*, a boldface **e**, or a capital E should be treated identically) and producing a transcription. We then identify each letter as a separate event, eliminating all nonletter characters such as numbers or punctuation. Finally, by compiling an event histogram and comparing it with the known distribution, we can determine a probability that the document was written in English. A similar process would treat each *word* as a separate event (eliminating words not found in a standard lexicon) and comparing event histograms with a standardized set such as the Brown histogram (Kuˇcera and Francis, 1967). The question of the comparative accuracy of these methods can be judged empirically.

The Burrows methods (Burrows, 1989; Burrows, 2003) for authorship attribution can be described in similar terms. After the document is canonicized, the document is partitioned into wordsevents. Of the words, most words (except for a chosen few function words) are eliminated. The remainder are collected in a histogram, and compared statisically to similar histograms collected from anchor documents. (The difference between the 1989 and 2003 methods is simply in the nature of the statistics performed.)

Even the “toutch” method can be so described; after canonicization, the event set of words, specifically, the number of words spelled “toutch.” If this set is nonempty, the document’s author is determined.

This framework also allows researchers both to focus on the important differences between methods and to mix and match techniques to achieve the best practical results. For example, (Juola and Baayen, 2003) describes two techniques based on crossentropy that differ only in their event models (words vs. letters). Presumably, the technique would also generalize to other event models (function words, morphemes, parts of speech), and and similarly other inference techniques would work on a variety of event models. It is to be hoped that from this separation, researchers can identify the best inference techniques and the best models in order to assemble a sufficiently powerful and accurate system.

5 Demonstration

The usefulness of this framework can be shown in a newly-developed user-level authorship attribution tool. This tool coordinates and combines (at this writing) four different technical approaches to authorship attribution (Burrows, 1989; Juola, 1997; Burrows, 2003; Kukushkina et al., 2000; Juola, 2003).

Written in Java, this program combines a simple GUI atop the threephase approach defined above. Users are able to select a set of sample documents (with labels for known authors) and a set of testing documents by unknown authors. The user is also able to select from a menu of event selection/preprocessing options and of technical inference mechanisms. Currently supported, for example, are three different choices — a vector of all the letters appearing in the sample/testing documents, a vector of all *words* so appearing, or a vector of only the fifty most common words/letters as previously selected, representing a restriction of the event model. Similarly, a variety of processing classes can be [have been] written to infer a similarity between two different vectors. Authorship of the test document can be assigned to (the author of) the most similar document.

Parties interested in seeing or using this program should contact the corresponding author.

6 Discussion and Future Work

The structure of the program lends itself easily to extension and modification; for example, the result of event processing is simply a Vector (Java class) of events. Similarly, similarity judgment is a function of the Processor class, which can be instantiated in a variety of different ways. At present, the Processor class is defined with a number of different methods [for example, crossEntDistance() and LZWDistance()]. A planned improvement is to simply define a calculateDistance() function as part of the Processor class. The Processor class, in turn, can be subclassed into various types, each of which calculates distance in a slightly different way.

Similarly, preprocessing can be handled by separate instantiations and subclasses. Even data input and output can be modularized and separated. As written, the program only reads files from a local disk, but a relatively easy modification would allow files to be read from a local disk or from the network (for instance, Web pages from a site such as Project Gutenberg or *literature.org*).

From a broader perspective, this program provides a uniform framework under which competing theories of authorship attribution can both be compared and combined (to their hopefully mutual benefit). It also form the basis of a simple user-friendly tool to allow users without special training to apply technologies for authorship attribution and to take advantage of new developments and methods as they become available. From a standpoint of practical epistemology, the existence of this tool should provide a starting point for improving the quality of authorship attribution as a forensic examination – by allowing the widespread use of the technology, and at the same time providing an easy method for testing and evaluating different approaches to determine the necessary empirical validation and limitations.

On the other hand, this tool is also clearly a "research-quality" prototype, and additional work will be needed to implement a wide variety of methods, to determine and implement additional features, to establish a sufficiently user-friendly interface. Even questions such as the preferred method of output — dendrograms? MDS subspace projections? Fixed attribution assignments as in the present system? — are in theory open to discussion and revision. It is hoped that the input of research and user groups such as the present meeting will help guide this development.

References

Burrows, J. (2003). Questions of authorships : Attribution and beyond. *Computers and the Humanities*, 37(1):5–32.

Burrows, J. F. (1989). 'an ocean where each kind. . . ' : Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(45):309–21.

Easson, G. (2002). The linguistic implications of shibboleths. In *Annual Meeting of the Canadian Linguistics Association*, Toronto, Canada.

Farringdon, J. M. (1996). *Analyzing for Authorship : A Guide to the Cusum Technique*. University of Wales Press, Cardiff.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Holmes, D. I. (1998). The evolution of stylometry in humanities computing. *Literary and Linguistic Computing*, 13(3):111–7.

Juola, P. (1997). What can we do with small corpora? Document categorization via crossentropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh, UK. Department of Artificial Intelligence, University of Edinburgh.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, 37(1):77–96.

Juola, P. (2004a). Adhoc authorship attribution competition. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden.

Juola, P. (2004b). On composership attribution. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteberg, Sweden.

Juola, P. and Baayen, H. (2003). A controlled corpus experiment in authorship attribution by cross entropy. In *Proceedings* of ACH/ALLC2003, Athens, GA.

Kukushkina, O. V., Polikarpov, A. A., and Khmelev, D. V. (2000). Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatii*, 37(2):96–198. Translated in "Problems of Information Transmission," pp. 172– 184.

Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University Press, Providence.

Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace Jovanovitch, Inc., Fort Worth, 3rd edition.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, IX:237–49.

Nerbonne, J. (2004). The data deluge. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteberg, Sweden. To appear in *Literary and Linguistic Computing*.

Stinson, D. R. (2002). *Cryptography: Theory and Practice*. Chapman & Hall/CRC, Boca Raton, 2nd edition.

Wellman, F. L. (1936). *The Art of Cross Examination*. MacMillan, New York, 4th edition.

Automatic Trend Detection and Visualization using the Trend Mining Framework (TMF)

Tobias Kalledat, School of Business and Economics of the Humboldt University in Berlin

1. Introduction

The documentation of historical Domain Knowledge about fields of activity usually takes place in the form of unstructured text -, picture -, audio- and video documents that are produced over longer time periods. Structured (e.g. relational data) and semi structured (e.g. HTML pages) and unstructured documents (e.g. texts) become distinguished with regard to the degree of an internal structure.

The usage of structured languages such as XML for tagging of textual data stands only at the beginning of its development path and is for historical documents therefore not to be found. Thus in the relevant literature is assumed that up to 80-90% of the electronically stored Knowledge is hidden in such unstructured sources [Tan99, Dörr00]. There are three implications most relevant:

a) The production of such documents rises over the time due to the distribution of Information Systems and their shared use, e.g. over the Internet. The availability of large sources of potentially interesting Knowledge becomes ubiquitous.

b) The usage of actual Knowledge becomes more and more a critical factor for competition of market participants. To adjust their product portfolio quickly it is necessary to adapt new ideas in a short development period, because consumers asking for product live cycles getting shorter and services getting closer to their individual needs. For members of the underlying processes studying lifelong gets more and more important and the flexibility of adapting new Domain Knowledge in a short time turns to one of the most important tasks.

c) Today the use of implicit Knowledge that is hidden in the huge amount of unstructured data is an approach that can be used because of the rapid development of powerful hardware that can handle such large data sources and the methods that were developed under the terms “Knowledge Discovery in Databases” (KDD) and "Data Mining" since the early 1990ies [Fayy96]. The Data Mining process allows discovering formerly unknown, potentially useful Knowledge Patterns out of various input data using techniques and methods of statistics, linguistics and data base research [Düsi99], [Biss99].

These are challenges for the educational sector as well as for the designers of Information Systems that support these processes of KDD. Therefore there is a market demand for turning the Knowledge Discovery Process (KDP) itself from an individual approach of a small number of specialists to a process that supports a large amount of "Domain Knowledge Researcher" in firms and organizations. Aim of the current research of the author is to find a methodology, which is able to support the KDP on large Technical Domain Corpora.

The Data Mining Process should be performed continuously. From the efficiency point of view it must be made sure, that Knowledge that was discovered in past time periods is be used as a basis for an adjustment of current decisions, which have relevance for the future.

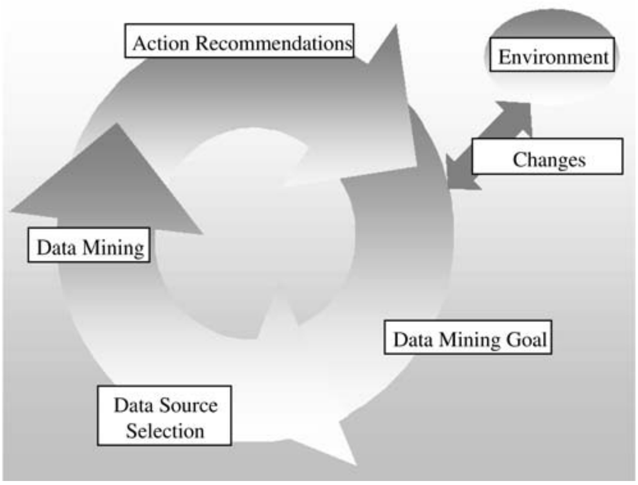


Fig. 1: Data Mining Management Cycle, translated from [Kall04]

Therefore a Data Mining Management Cycle was being proposed in [Kall04], see Fig. 1. While analyzing large amounts of textual data, for the Domain Knowledge Researcher two problems result: On the one hand the pure amount of unstructured historical and present data represents a substantial entrance barrier. On the other hand unstructured data themselves cannot be automatically processed easily. A substantial realization gain is to be expected, if methods are found to open and evaluating the mentioned unstructured sources of information.

The advantage of an evaluation of historical text documents in the opposite to interviewing time witnesses exists in the reflection of the historical reality free of subjective information distortions. On this assumption, the provision of information problem reduces to the procurement of suitable text documents. Expenditures for the determination of time witnesses and for the execution and evaluation of interviews can be minimized thereby. The analysis of electronic sources gives also the opportunity to transform semantic structures of Knowledge Domains in an Information Technology (IT)-based representation that allows documenting and sharing Knowledge more easily.

To support individuals in the KDP is economically interesting. Potentially expensive manual work can be substituted by automatically working Information Technology driven solutions. Approaches in this support are basing mostly on methods that working on each text file itself for clustering, tagging or classifying purposes [Lend98, p.130], [Moen00]. The objective mostly is to support information retrieval or later querying against a mass of these text files that were proceeded the same way. Most of the used procedures do not consider the time dimension.

Technical Domains, e.g. Information Technology, do have special qualities, which make it necessary to configure the methods of research to the needs of the research aim. Such corpora must be handled different to "usual" corpora. For example, product names, programming languages and other proper names must be kept during all analysis steps. Pure linguistic approaches therefore not applicable. Therefore instead of "Word" the term "Phrase" is used from now which covers a wider range of alphanumeric sign combinations, e.g. product names and technical norms.

For a Domain Knowledge Researcher it is important to know: How does the semantic of Phrases change over the time? Which topics are growing, falling and what is the semantically basis of the Domain? For distinguishing between these clusters, rules for significant decisions are needed. An important challenge for research is to define methods, which can extract significant pattern and track time dependent changes.

The main specifics of the proposed methodology are:

1) Proposing a top down approach for pre-filtering such patterns, that are worth for further analysis using appropriate Meta Data Measures of corpora

2) Proposing the use of time dependent Ontology for tracking Trends and covering semantic changes

3) Suggesting a visualization concept, which visualizes the Domain Knowledge, that is represented by a corpus

4) Combining the top down Meta Data Measure approach with the Ontology based semantic modelling using an OLAP-based intuitive navigation concept

The objective of this paper is to propose and evaluate appropriate methods for Automatic Detection, Classification and Visualization of Trends in large technical focused Domain Corpora. Chapter 2 introduces common Approaches for pattern recognition in textual data; chapter 3 is dealing with the proposed methodical approach of the "Trend Mining Framework". It's components and the proposed Trend Mining Process are introduced in chapter 4. The paper closes with conclusions and outlook in chapter 5.

2. Approaches for pattern recognition in textual data

Since the 1990-ies under the term Data Mining methods were developed, which make it possible to recognize unknown structures in data and derive from it action-relevant and economical useful Knowledge [Codd93] . These methods are based on classical statistic procedures as well as methods of adjacent research fields and were adapted for the employment on appropriate data.

Methods for the investigation of unstructured data, e.g. large text corpora or speeches, usually subsumed under the headline Computer Linguistics or Content Analysis. As a special research field Text Mining was developed for the computer-based analysis of unstructured textual data. In the Mid-1990ies the research activities get pushed by carrying out the "*Topic Detection and Tracking*" (TDT) task by a few research organizations and the U.S. government. Subject of this research is event-based organization of broadcast news [Alla02a], [Alla02b] . The main research task was divided into the following sub-tasks: *Story Segmentation, First Story Detection, Cluster Detection, Story or Topic Tracking* and *Story Link Detection.* Most of the used methods are bottom up approaches that analyzing text corpora word-by-word or sentence-by-sentence and using clustering and tagging techniques [Spil02]. Other methods are more statistically based and working with Features, e.g. corpus wide measures or Vector Space Models which represent sentences or whole stories. Applications based on these methods are realized, e.g. for *Automatic News Summarizing, Document Clustering and Patent Mining.*

The classical linguistic text analysis has a long developmental history, which reaches back up to time periods of the middle of the last century. An important influence on linguistics Chomsky had, who characterized the research in the middle of the last century. He especially criticized the use of the Corpus Linguistics approach for learning more about the language itself, because all analysed corpora can only be a subset of the whole variety of the language. With this "generative transformational grammar" he revolutionized linguistics and became the most important theorist of his branch. Modern research approaches are using statistical methods for quantitative analyses of text corpora, e.g. for semantic similarity checks of document sets. The linguistic methods therefore can be differed into two main directions of research activity:

(1) The predominant quantitative approach, that uses Meta Data Measures of text corpora, e.g. term or word frequency for evaluation and comparison of different text sources [Atte71]. Some researchers are assuming, that all the rules are inside the used language, that it is worth analysing real life data to learn more about languages. Purposes for this Corpus Linguistics approach e.g. comparisons between authors and their styles or detecting, whether a text belongs to author A or author B.

(2) The more qualitative approach of descriptive grammars [Lend98, p. 106] , that makes use of interpretation techniques and is working with thesauri or special grammatical algorithms for word analysing purposes (e.g. stemma finding). Goal is describing languages with rules and building a computable basis, e.g. for Machine Translation solutions.

It can be observed that most of the methods for text mining are bottom up approaches, coming from the smallest unit of textual data, a word or n-gram (only a few letters) and generalizing the pattern, which were found. For the tracking of Trends in Technical Domain Corpora over time these known bottom up procedures are having limitations:

i. There are performance issues when real large corpora is analysed.

ii. The patterns found are based on generalized results of multi parametric algorithms, which means that there is to be expected a biased result due to the multiplication of error terms.

iii. Linguistic approaches are not appropriate, because technical information is not covered or is destroyed during stemma finding processes.

iv. The generation of action recommendations is not transparent to "normal" users.

To support the process of Automatic Detection, Classification and Visualization of Trends in large Technical Domain Corpora instead of using usual bottom up procedures a promising approach is to combine classical Meta Data Measure oriented analysis in the first step with a projection step of found patterns into the detail layer of Phrases, in order to overcome this lack of research. Analysing appropriate Meta Data in the first step should reduce the complexity for pattern recognition dramatically. After this, the relations between interesting Phrases can be analysed more focused. Appropriate methods for modelling hierarchical structures of elements, e.g. Phrases, are known under the term Ontology. These are semantic concepts modelling Knowledge Domains by the use of directed graphs, which can show relations between elements of Domain Corpora. Based on such concepts deeper analyses, e.g. the use of classical Data Mining techniques is possible in later steps. The proposed methodology is introduced in the following chapters.

3. Trend Mining Framework

The TMF is a proposed methodology and also a process of Text Mining, which makes it possible to extract time-related Domain Knowledge based on unstructured textual data semi automatically. To establish a framework that allows exploring and analysing large Technical Domain Corpora in an intuitively interactive way is the main goal of developing the TMF approach. For this, methods were evaluated, which allow measuring the quantitative characteristics of a time tracked Domain Corpus.

A basis assumption is, that the frequency of Phrases is positive correlated with their importance within a corpus (no articles et cetera, but nouns and names). Based upon this assumption possible Meta Data Measure candidates can be defined, e.g.:

I. In [Crou90] the "Term Frequency Inverse Document Frequency" (TF-IDF) is used, which is defined as the number of times a Phrase appears in a document multiplied by a monotone function of the inverse number of documents in which the Phrase appears. In difference to the original source here the former defined term "Phrase" was used instead of the term "word".

II. "Type-Token Ratio" (TTR) is defined as the ratio of different Phrases to their number of occurrence in a corpus [Lend98].

III. The "Phrase Repetition Quota" (PRQ) or Frequency is the ratio between all different Phrases to all Phrases occurring in a corpus.

Initially, the PRQ was used as Meta Data Measure in a prototypic realization of the Trend Mining

Process on a large corpus. The TF-IDF, TTR and monotone transformations of them as well as further candidates seem to be worth for testing in later realizations. It has to be taken under consideration, that only a few (short) words or Phrases are occurring often in a corpus. More important words or Phrases are longer, but rare [Zipf49]. A usual method is to filter very frequent, but not meaningful "stop words" to focus the analysis on semantically more important Phrases. In the current investigation a short stop-word list of some very frequent terms was used for filtering.

As an example, the historical development of the Domain "Information Technology" was analysed using the TMF. By the use of the TMF methodology it was possible to identify temporal semantic developments and to differentiate these regarding its persistence characteristics in short living Hypes and long-term Trends and to describe their development paths while keeping the technical Peculiarities of the sources. Extracting corpus Meta Data and transforming it into an appropriate detail level using a Domain specific Ontology do this. The TMF consists of

1. **Format Parsers**: for separation of content and formatting as well as converters to ASCII
2. **Analytical Parsers**: for determination of the Meta Data of the corpus and Feature Extraction, e.g. Phrase frequency
3. **Data Warehouse**: for the storage, e.g. of intermediate results and for a semi-automatic assignment of Phrases to an Ontology out of the Domain Corpus
4. **Statistical Tools**: for analyses and tests
5. **Graphical component**: for visualization and navigation

In the next chapter the proposed Trend Mining Process using the TMF is described in detail.

4. Trend Mining Process

In the TMF process the main task of Automatic Detection, Classification and Visualization of Trends in large Technical Domain Corpora is divided into a few sub tasks. The tasks are performed by the different components, mentioned earlier.

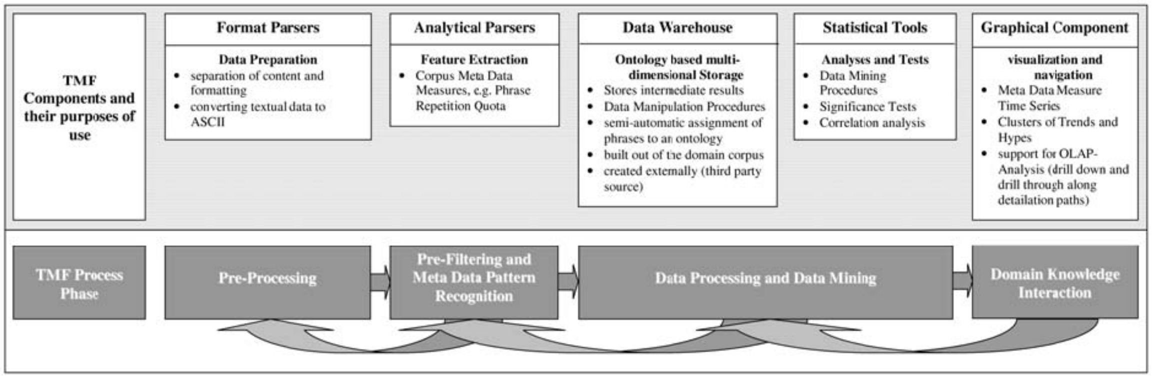

TMF Components and their purposes of use
Format Parsers
Data Preparation
• separation of content and formatting
• converting textual data to ASCII
Analytical Parsers
Feature Extraction
• Corpus Meta Data Measures, e.g. Phrase Repetition Quota
Data Warehouse
Ontology based multi-dimensional Storage
• Stores intermediate results
• Data Manipulation Procedures
• semi-automatic assignment of phrases to an ontology
• built out of the domain corpus
• created externally (third party source)
Statistical Tools
Analyses and Tests
• Data Mining Procedures
• Significance Tests
• Correlation analysis
Graphical Component
visualization and navigation
• Meta Data Measure Time Series
• Clusters of Trends and Hypes
• support for OLAP-Analysis (drill down and drill through along detailation paths)
TMF Process Phase
Pre-Processing
Pre-Filtering and Meta Data Pattern Recognition
Data Processing and Data Mining
Domain Knowledge Interaction

Fig. 2: Components of the TMF and their use in the Trend Mining Process

In Fig. 2 the Components of the TMF and their use in the Trend Mining Process are shown. The Trend Mining Process consists of the phases: *Pre-Processing, Pre-Filtering and Meta Data Pattern Recognition, Data Processing and Data Mining as well as Domain Knowledge Interaction*. Prototypically the TMF was used to analyse a large technical orientated corpus of the German issue of the weekly magazine “Computerwoche” (engl.: “Computer Week”) of the last 29 years, starting in 1975 [Kall03]. For this paper the articles of the year 2003 added to the corpus and the results were updated.

In the **Pre-Processing Phase** the 132.406 articles were converted to pure ASCII files by the use of Format Parsers. In the phase of Pre-Filtering and Meta Data Pattern Recognition, Corpus Meta Data Measures were extracted for all available articles of each year of publication. All articles of one year were analysed together at the same time. It was counted 29,777,462 Phrases for the whole corpus and 2,941,680 distinct Phrases, which means an average PRQ of 10.12. The PRQ was used for the Top Down Approach of Pattern Recognition. The Graph of the PRQ is shown in Fig. 3:

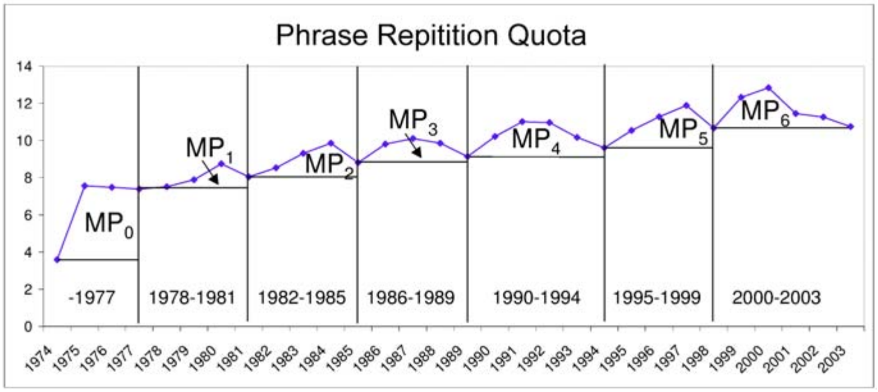


Fig. 3: Graph of Phrase Repetition Quota for each year of publication

The PRQ is rising on the average and periods are visible, which are divided by turn points of the PRQ (rising falling rising). The areas, marked as M_{P0}… M_{P6} are patterns that are worth to be analysed, because they are representing Phrases, which do have a Frequency that is higher than the average PRQ. The Frequencies where normalized that the specific corpus length of each year issues does not influence their comparability. After this **Phase of Meta Data Pattern Recognition** the articles of the corpus were time based segmented according to the periods, which were recognized. All further analyses are based on the segmented corpus. The phenomenon of an extensive PRQ being carried out via a restricted temporal period is understood by a “Hype” in the Information Technology in the following. Unlike such a thematic excess concepts, which learned, an extensive mentioning in several periods can exist. Such concepts are described below as Trends. Hypes are rather "nine days' wonders" after this interpretation, without considerable durable meaning, “Trends”, on the other hand, characterize the Information Technology in the long run. Phrases, which have a Frequency higher than the lowest PRQ period limit, defining periods of “Trend” or “Hype” Phrases, which occurring more often in the corpus than other Phrases in the same period. Using this filter criterion, the number of Phrases was reduced by 16%. Thus, new time based corpora is analysed by considering results from former corpora also a basic learning capability is present.

From the viewpoint of the Set Theory, during a first step in the Phase of **Data Processing and Data Mining** it is possible to subsume the whole corpus as $MV_{CW1974\_2003} = M_{CW1974} \cup M_{CW1975} \cup \ldots M_{CW2003}$, which is equal with the Union Set of all articles. In the following steps, the previously built periodical segments of the corpus are clustered in a few Sub Sets as a preparation for further analysis. The average set of the Phrases of all years is $MK_{CW1974\_2003} = M_{CW1974} \cap M_{CW1975} \cap \ldots M_{CW2003}$ (approx. 50% of the corpus Phrases). $MK_{CW1974\_2003} = MK_L \cup MK_D$, where $MK_{L,}$ contains Phrases that are typical for the language of which the corpus consists and MK_D, which consists of Phrases that are representing the constant basis of the Knowledge Domain. For the last period found (1999-2003), after an additional subtracting of Phrases, which belonging to MK_{CW1974_2003}, the number of Phrases was reduced by 82% until this step. In order to differ between the been left Phrases the usage of Ontology based semantically concepts are needed. For this analysis a prototypical Ontology was build manually out of the corpus. The Phrases were assigned to dimensions, e.g. *Vendor (German: Anbieter)*, *Programming Language*, et cetera. The built Ontology is a simple directed graph representation of the semantic of the Phrases of the corpus. In this step, also existing externally defined and more complex Ontology’s can be integrated instead.

For the Phase of **Domain Knowledge Interaction** the previously prepared and clustered corpus data is presented to the user in a way, which allows not only visualization, but also an interactive analysis of the data. The philosophy of the visualization component as a core part of the TMF is based on the metaphor of a “Trend Landscape”. Two-dimensional concepts, e.g. "ThemeRiver" [Havr02] at which selected Frequencies are represented in a kind of topic flow in the course of time, are limited concepts conditionally in their representation and interaction ability.

A: Schematic View in Fig. 4 shows, how the Meta Data (M) of the text corpus (e.g. Frequency or monotone Transformations of it), which is represented in the two-dimensional Phrase/time layer, in the projection step (P) into the three-dimensional detail dimension is to be projected. According the geographical metaphor the Meta Data become represented as “River” in the Trend Landscape. The Hypes (H) and Trends (T) defined before are rising as clusters, the so-called Dimensional Mountains over the set of the Phrases that occurring in all periods (MK_L and MK_D). The methodology of the Trend Landscape also supports the idea of the OLAP analysis methods “Slice and dice” as well as ”Drill down and Drill through” and navigating along detailing paths (e.g. “Market participant” “Vendor” „Netscape“) using the Ontology, which was built, based on the Domain Corpus. In every step of interaction it is possible to disaggregate or aggregate the view on the data. This capability is similar to usual Data Warehouse Tools. In B: Clustered Segment View (Dimension Level) Dimensions that are part of MK_D are shown clustered in the periodical patterns, which were introduced earlier. The Rank of Phrases was transformed that a higher value represents a higher Frequency of the Phrases.

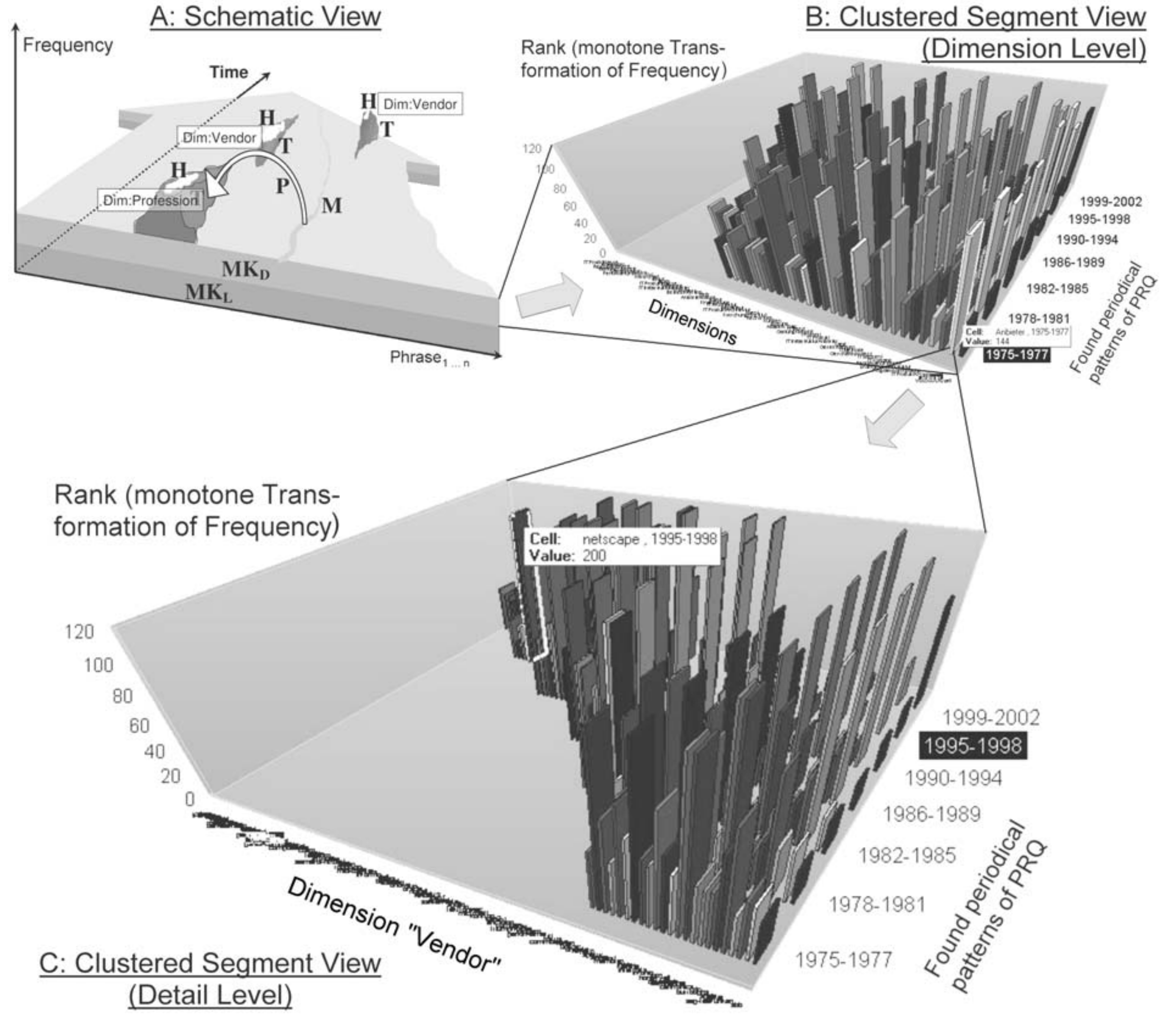


Fig. 4: "Trend Landscape” Views in the process of Domain Knowledge Interaction

The “Drill down” of the Dimension “Vendor” is shown in C: Clustered Segment View (Detail Level). It can be seen that some of the Phrases do not occurring in every periodical patterns, e.g. “Netscape”, which is present only from 1995-1998 as a Phrase with a higher Frequency than the average PRQ. Here it is possible to distinguish between a Hype Phrases (single occurrence) and a Trend Phrases (multiple occurrence).

5. Conclusions and Outlook

The analysis of a large Technical Domain Corpus was properly supported by the TMF. Meta Data Patterns of PRQ were used to segment the Phrases of the corpus due to their persistence character over time. The built segments can analysed separately, according to the aim that the Domain Knowledge Researcher wants to reach. Using Ontology's, it is possible to assign each Phrase to a dimensional structure, which allows navigating through different views of the prepared Domain Knowledge representation by intuitive OLAP navigational concepts. Thus, the TMF concept is open, there are a few degrees of freedom for adapting this procedure regarding individual research needs by: a) Using different Meta Data Measures. b) Integrating various mathematical or statistical procedures into the Projection step (P). c) The rules for building clusters (in the actual example: Trends and Hypes) can be free defined. To track semantically changes, e.g. the changing of the semantics of the Phrase "Mailbox" over the time, instead of the current used manually built semantically Ontology the use of similarity measures is appropriate. In [Senel04] the Term Document Frequency (TDF) as a candidate discriminator measure is proposed instead of a hard computable cosine of all combinations within a term vector space model. They found, that terms, which appearing in 1 to 10% of the documents are good discriminators. Based on the assumption that Phrases, which are good discriminators for documents, also good discriminators for Semantic Dimensions, in further realizations additional measures for similarity are planned to be used in future.

Sources

[Alla02a] Allen, James : Introduction to Topic Detection and Tracking . Hrsg.: Allen, James : Topic Detection and Tracking: Event-based Information Organization . Massachusetts , Kluwer Academic Publishers , 2002

[Alla02b] Allen, James; Lavrenko, Victor; Swan, Russel : Explorations within Topic Tracking and Detection . Hrsg.: Allen, James : Topic Detection and Tracking: Event-based In-formation Organization . Massachusetts , Kluwer Academic Publishers , 2002

[Atte71] Atteslander, P. : Methoden der empirischen Sozialforschung (Methods of empirical social research) . 2. Auflage , Berlin , de Gruyter , 1971

[Biss99] Bissantz, Nicolas : Aktive Managementinformation und Data Mining: Neuere Methoden und Ansätze . Hrsg.: Chamoni, Peter; Gluchowski, Peter : Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining . 2. Auflage , Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999 ISBN 3-540-65843-2.

[Codd93] Codd, E.F.; Codd, S.B.; Sally, C.T. : Providing OLAP (on-line analytical processing) to user-analysts – an IT mandat. White Paper . E.F. Codd & Associates , 1993

[Crou90] Crouch, C. J.: An approach to the automatic construction of global thesauri. Information

Processing and Management. 1990, 26, p. 629-640

[Dörr00] Dörre, J.; Gerstl, P.; Seiffert, R. : Text Mining . Hrsg.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D. : Handbuch Data Mining im Marketing . Braunschweig , Vieweg , 2000

[Düsi99] Düsing, Joachim : Knowledge Discovery in Databases und Data Mining . Hrsg.: Chamoni, Peter; Gluchowski, Peter : Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining . 2. Auflage , Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio , Springer , 1999 ISBN 3-540-65843-2. .

[Fayy96] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. : From data mining to knowledge discovery: an overview . Hrsg.: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. : Advances in knowledge discovery and data mining . Menlo Park (Califor-nia) , 1996 S.p. 1-34 ,

[Havr02] Havre, S.; Hetzler, E.; Whitney, P.; Nowell, L. : ThemeRiver: Visualizing thematic changes in large document collections . IEEE Transactions on Visualization and Com-puter Graphics . 2002 , *8(1)*, Jan – Mar 2002

[Kall03] Kalledat, T. : Separation of long-term constant elements in the field of information technology from short existing trends based on unstructured data . Hrsg.: Viehweger, B. : Perspectives in Business Informatics Research, Proceedings of the BIR-2003-Conference . Aachen, Shaker Verlag , 2003 S. 167-183 ,

[Kall04] Kalledat, T. : Perspektiven der Nutzung von Data-Mining-Technologien in der Energieversorgungswirtschaft . ew-Elektrizitätswirtschaft – Das Magazin für die Energiewirtschaft . 2004 , *7*, S. 48-53 , 1619-5795-D9785D.

[Lend98] Lenders, Winfried; Willée, Gerd : Linguistische Datenverarbeitung . 2. Auflage , Opladen/ Wiesbaden , Westdeutscher Verlag , 1998

[Moen00] Moens, Marie-Francine : Automatic Indexing and Abstracting of Document Texts. Massachusetts , Kluwer Academic Publishers , 2000

[Senel04] Senellart, Pierre P.; Blondel, Vincent, D. : Automatic Discovery of Similar Words . Hrsg.: Berry, M. : Survey of Text Mining: Clustering, Classification and Retrieval . New York , Springer , 2004

[Spil02] Spiliopoulou, M.; Winkler, K. : Text Mining auf Handelsregistereinträgen: Der SAS Enterprise Miner im Einsatz . Hrsg.: Klaus D. Wilde, Hajo Hippner, Melanie Merzenich : Data Mining: Mehr Gewinn aus Ihren Kundendaten . Düsseldorf , Verlagsgruppe Handelsblatt , S.S. 117-124 ,

[Tan99] Tan, A. -H. : Text Mining: The State of the Art and the Challenges . Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases . Peking , 1999 S.S. 65-70 ,

[Zipf49] Zipf, G. K. : Human Behavior and The Principle of Least Effort . Cambridge, Mass , Addison-Wesley , 1949

In pursuit of radiance: Report on an interface developed for the Piers Plowman Electronic Archive

Eugene W. Lyman, Boston University

The age of archives has not ended by any means, but the age of tools has clearly begun. – ***Stephen Ramsay***

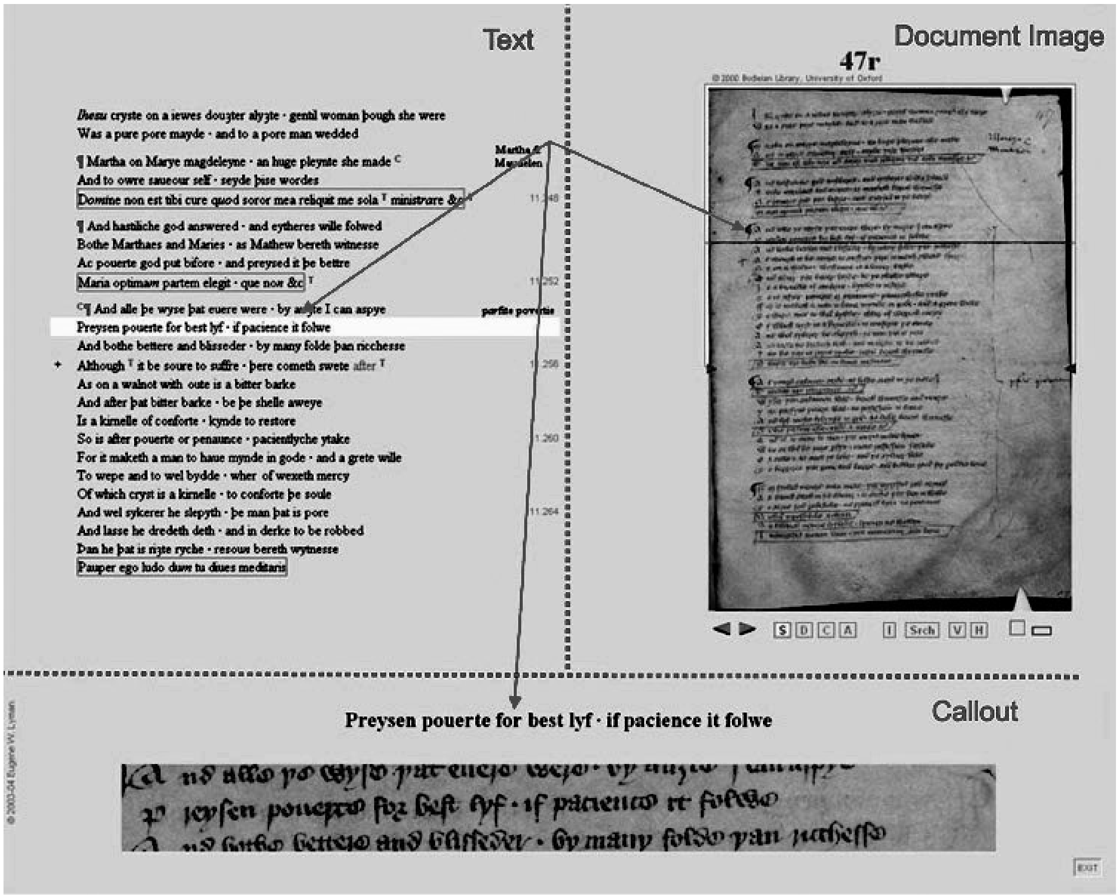
Interface enables and reflects the reader's active presence, it is the environment where readers live and move and have their being in digital simulations. ***– Jerome McGann***

After nearly two decades devoted to the elaboration of text encoding schemes and to the actual encoding of large volumes of text, interface has become a focus of pointed attention. The timing of this notice has much less to do with interface's relative importance in the delivery of truly useable electronic texts than with what we might think of as a simplified "just in time" approach collectively adopted by the pioneers in the field. First we needed a collection of texts encoded in ways that facilitate serious interrogation. Now we need the lens through which we can see to begin our interactive questioning. The development of such a lens is by any measure a non-trivial project. As Jerome McGann has noted, "designing interfaces that are at once stable and flexible, stimulating as well as clear, is one of the two most demanding tasks -- in both senses of 'demanding' -- now facing the scholar who means to work with digital tools."

The production of electronic editions of individual manuscripts of the Middle English poem Piers Plowman by the Piers Plowman Electronic Archive aptly illustrates both the challenge of interface design – especially as it pertains to a multimedia projects involving texts and images – and its current status as a focus of needed attention. At the Archive's inception in the early 1990's its principal architect, Hoyt Duggan, noted that the work of his group in encoding and editing the texts of the 54 surviving Piers manuscripts would provide scholars the basis for future interrogation "in ways unimagined by its creators." At the outset, limitations in available hardware and display software reinforced the PPEA's commitment to the platform independent standards of TEI markup. Like other archival projects launched in this period, the PPEA began its work with an expectation that the development of an interface suited to the complexity of its work would follow in due course.

Two years ago, after I had been enlisted to edit a manuscript of the A version of Piers, I set out to develop such a piece of software. My presentation focuses on the issues that I encountered in creating an interface that seeks to expose the work of the PPEA's editors to full advantage and that equips readers with tools that encourage sophisticated interrogation of the text and document images supplied in their editions.

The interface that has emerged from this effort provides for close coordination of text and digital image as well as for the provision of constant visual cues to indicate a reader's location within a document. Figure 1 illustrates both of these principles as they are applied to viewing a folio leaf of the text (labels, arrows, and dotted lines are for explanatory purposes and are not a part of the display).



Ihesu cryste on a iewes douȝter alyȝte · gentil woman þough she were
Was a pure pore mayde · and to a pore man wedded

¶ Martha on Marye magdeleyne · an huge pleynte she made [C]
And to owre saueour self · seyde þise wordes
Domine non est tibi cure quod soror mea reliquit me sola [T] ministrare &c

¶ And hastiliche god answered · and eytheres wille folwed
Bothe Marthaes and Maries · as Mathew bereth witnesse
Ac pouerte god put bifore · and preysed it þe bettre
Maria optimam partem elegit · que non &c [T]

[C]¶ And alle þe wyse þat euere were · by augte I can aspye
Preysen pouerte for best lyf · if pacience it folwe
And bothe bettere and blisseder · by many folde þan ricchesse
+ Although [T] it be soure to suffre · þere cometh swete after [T]
As on a walnot with oute is a bitter barke
And after þat bitter barke · be þe shelle aweye
Is a kirnelle of conforte · kynde to restore
So is after pouerte or penaunce · pacientlyche ytake
For it maketh a man to haue mynde in gode · and a grete wille
To wepe and to wel bydde · wher of wexeth mercy
Of which cryst is a kirnelle · to conforte þe soule
And wel sykerer he slepyth · þe man þat is pore
And lasse he dredeth deth · and in derke to be robbed
Þan he þat is riȝte ryche · resoun bereth wytnesse
Pauper ego ludo dum tu diues meditaris

Preysen pouerte for best lyf · if pacience it folwe

Figure 1.

As a reader mouses over either the text or document image panels, the area receiving focus is indicated by the concurrent highlighting of a line of text along with the display of a red line beneath the corresponding passage in the document image. The display is dynamic: as the cursor position is moved across either panel, the highlighting and underscore shift to reflect the change in focus. While these features keep the reader oriented as to the position of the selected text within the document leaf, a callout panel displays an enlarged section of the document image as well as the corresponding textual transcription – thereby responding the reader's indicated attention with a display of the appropriate portion of the document at a level of screen resolution that promotes easy legibility. The contents of the callout panel are thus fully coordinated with the visual cueing that takes place in the text and document image panels.

Although the display described in the foregoing paragraph is not an analytical tool such as we might typically imagine, it provides substantial assistance to the visual processing of text and image. It does so by being a mechanism that simultaneously responds to a reader's focus of attention while guiding it to and around the specific portion of the document that he or she has identified for closer inspection. Such a display aspires to be a tool in the very broad and fundamental sense: enabling and reflecting the reader's active presence – to borrow from Jerome McGann's apt description of interface in one of the quotations that opens this abstract.

The interface that I have developed also presents more straightforwardly analytic tools, equipping readers with the means to undertake the active examination of a source document's text – as well as its imaged representation. Readers can choose, for example, to enlarge or apply color filters to specific portions of document images without having to toggle between multiple windows or frames. The dual display of the enlarged portion of the document alongside an image of the entire leaf that is marked to indicate the location of the enlargement serves again to preserve the reader's visual orientation within the document. Any one who has attempted to search for a detail by scrolling through an image that has been enlarged past the confines of a computer display will understand the benefit that this preservation of visual context provides.

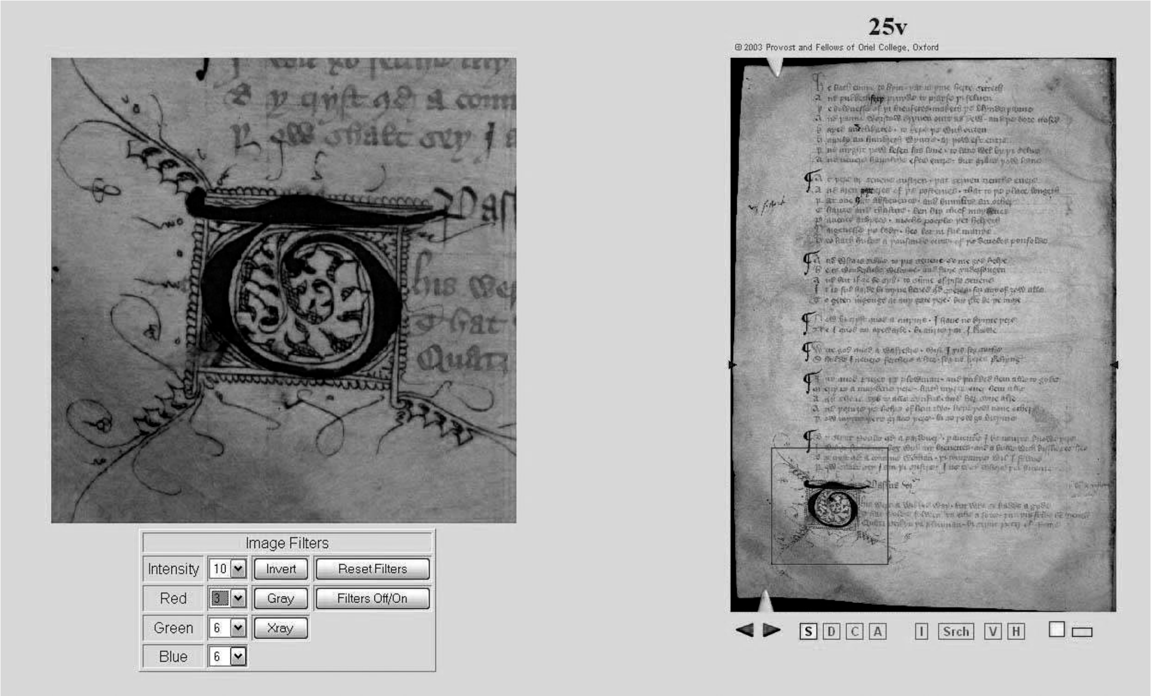


Image Filters			
Intensity	10	Invert	Reset Filters
Red	3	Gray	Filters Off/On
Green	6	Xray	
Blue	6		

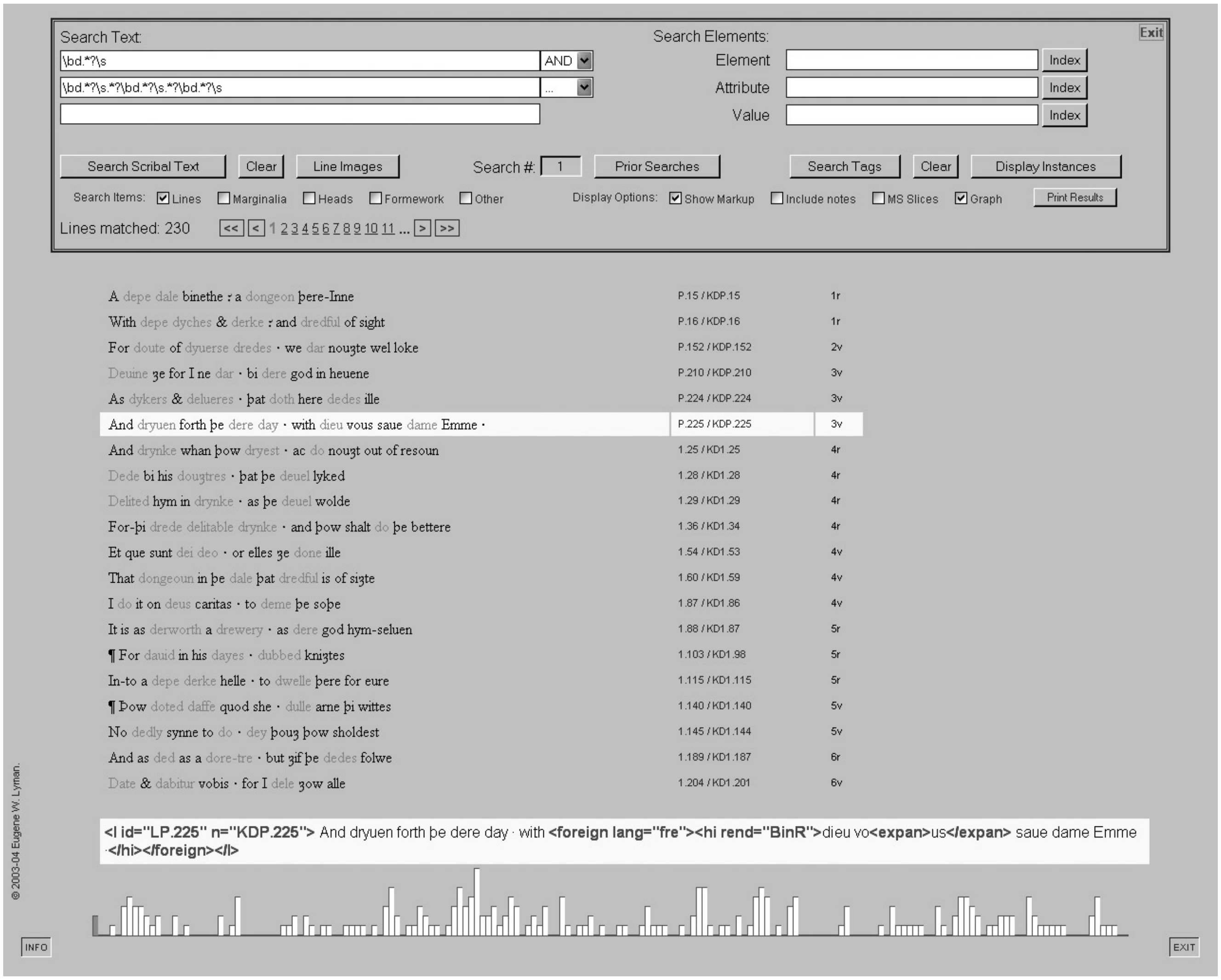
25v

© 2003 Provost and Fellows of Oriel College, Oxford

Figure 2.

The interface also provides readers with the means to conduct complex Boolean searches of the document's text and XML markup using specific words and phrases as well as broadly generalizable regular expressions. The results of these searches are presented as a concordance-on-the-fly with found words given emphasis by being rendered in red. Clicking on an entry in this concordance brings about the display to the folio in which the found line occurs. Distributions of items found by any search across the all manuscript folios may also be displayed in the form of graphic representations available in the viewer's search or document display modes and in the form of numeric tables that can be extracted for use with specialized statistical programs.

Figure 3 illustrates the results of a search in which regular expressions have been used to find lines that contain at least three words alliterating on an initial /d/. The bottom of the figure displays a graphic representation of the distribution of the found lines within the manuscript. Also illustrated in this figure is the interface's capacity to display the text's underlying xml markup.



Search Text:
\bd.*?\s AND
\bd.*?\s.*?\bd.*?\s.*?\bd.*?\s ...

Search Elements:
Element
Attribute
Value

Search #: 1

Search Items: ☑ Lines ☐ Marginalia ☐ Heads ☐ Formework ☐ Other

Display Options: ☑ Show Markup ☐ Include notes ☐ MS Slices ☑ Graph

Lines matched: 230

A depe dale binethe ∶ a dongeon þere-Inne	P.15 / KDP.15	1r
With depe dyches & derke ∶ and dredful of sight	P.16 / KDP.16	1r
For doute of dyuerse dredes · we dar nouȝte wel loke	P.152 / KDP.152	2v
Deuine ȝe for I ne dar · bi dere god in heuene	P.210 / KDP.210	3v
As dykers & delueres · þat doth here dedes ille	P.224 / KDP.224	3v
And dryuen forth þe dere day · with dieu vous saue dame Emme ·	P.225 / KDP.225	3v
And drynke whan þow dryest · ac do nouȝt out of resoun	1.25 / KD1.25	4r
Dede bi his douȝtres · þat þe deuel lyked	1.28 / KD1.28	4r
Delited hym in drynke · as þe deuel wolde	1.29 / KD1.29	4r
For-þi drede delitable drynke · and þow shalt do þe bettere	1.36 / KD1.34	4r
Et que sunt dei deo · or elles ȝe done ille	1.54 / KD1.53	4v
That dongeoun in þe dale þat dredful is of siȝte	1.60 / KD1.59	4v
I do it on deus caritas · to deme þe soþe	1.87 / KD1.86	4v
It is as derworth a drewery · as dere god hym-seluen	1.88 / KD1.87	5r
¶ For dauid in his dayes · dubbed kniȝtes	1.103 / KD1.98	5r
In-to a depe derke helle · to dwelle þere for eure	1.115 / KD1.115	5r
¶ Þow doted daffe quod she · dulle arne þi wittes	1.140 / KD1.140	5v
No dedly synne to do · dey þouȝ þow sholdest	1.145 / KD1.144	5v
And as ded as a dore-tre · but ȝif þe dedes folwe	1.189 / KD1.187	6r
Date & dabitur vobis · for I dele ȝow alle	1.204 / KD1.201	6v

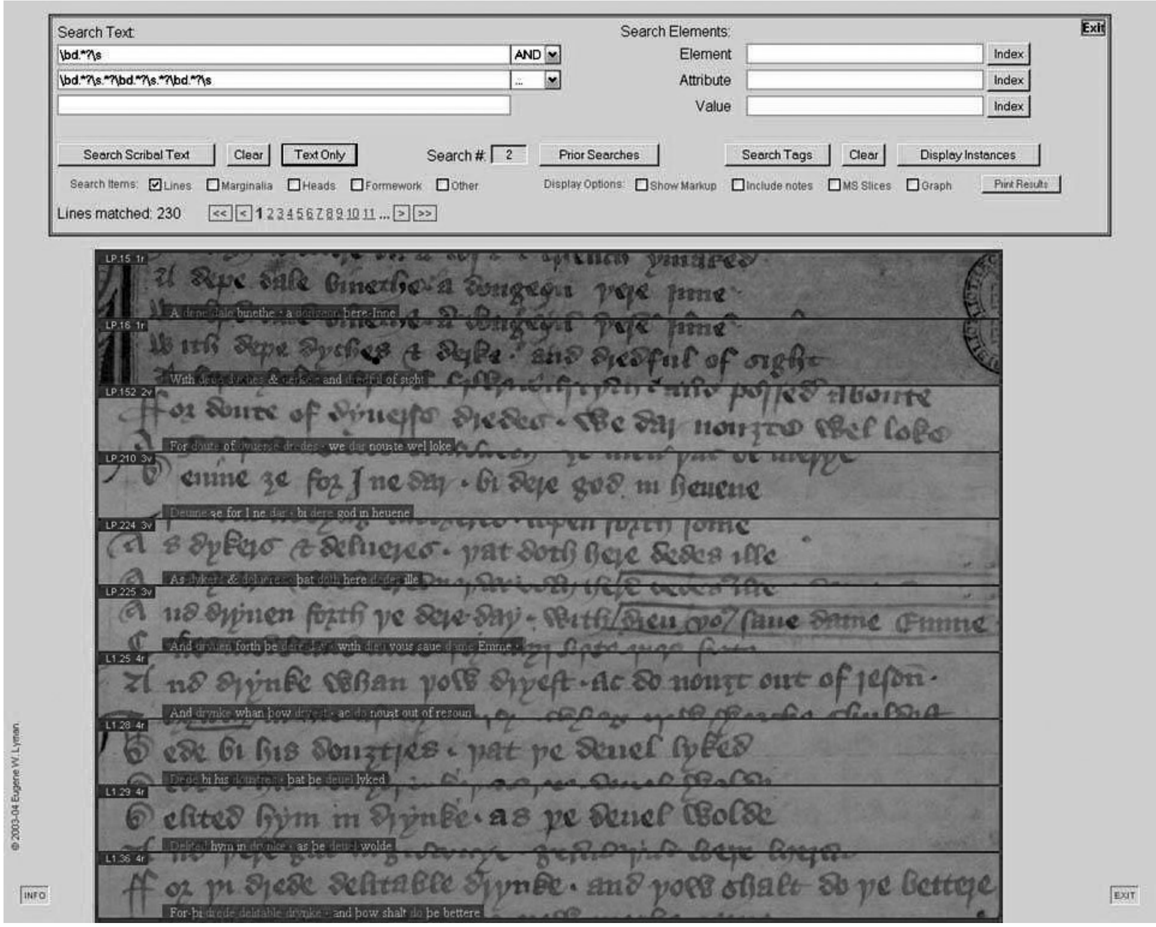
```
<l id="LP.225" n="KDP.225"> And dryuen forth þe dere day · with <foreign lang="fre"><hi rend="BinR">dieu vo<expan>us</expan> saue dame Emme
·</hi></foreign></l>
```

© 2003-04 Eugene W. Lyman.

Figure 3.

A final illustration of some of the tools incorporated in this interface returns to the guiding principle of seeking to maintain close coupling of transcription text and document image. Figure 1 demonstrates how this coupling eases the work that reader would do to correlate portions of the text and image if the transcription and the document images were opened in disconnected windows (or frames). Figure 4 carries the interface's coordination of text and image to a new level. It illustrates the same search undertaken in Figure 3, but with the display of the relevant line images directly drawn from the manuscript images that are a part of the edition instead of lines from the transcription.

The resulting "line image concordance" allows a reader to examine a collection of lines lifted from the their original context in folio leaves. Using this concordance feature, a reader is able to make direct comparisons of manuscript features that would normally place a heavy tax on the visual memory of any one who wished to make them. Owing to the viewer's capability to search on xml markup for "non-textual" document characteristics that have been noted by the PPEA's editors, features such as changes in scribal hand, manuscript damage, etc. may be compared with relative ease, thereby extending the interface's value for codicological and paleographical inquiry.



Search Text: \bd.*?\s AND
\bd.*?\s.*?\bd.*?\s.*?\bd.*?\s

Search #: 2

Search Items: ☑ Lines ☐ Marginalia ☐ Heads ☐ Formework ☐ Other

Display Options: ☐ Show Markup ☐ Include notes ☐ MS Slices ☐ Graph

Lines matched: 230

LP.15 1r
A depe dale binethe · a dongeon þere-Inne

LP.16 1r
With depe dyches & derke · and dredful of sight

LP.152 2v
For doute of dyuerse dredes · we dar nouȝte wel loke

LP.210 3v
Demme ȝe for I ne dar · bi dere god in heuene

LP.224 3v
As dykers & deluers · þat doth here dedes ille

LP.225 3v
And dryuen forth þe dere-day · with dieu vous saue dame Emme ·

L1.25 4r
And drynke whan þow dryest · ac do nouȝt out of resoun

L1.28 4r
Dede bi his douȝtres · þat þe deuel lyked

L1.29 4r
Delited hym in drynke · as þe deuel wolde

L1.36 4r
For-þi drede delitable drynke · and þow shalt do þe bettere

Figure 4.

In addition to demonstrating these and other features of the current version of this interface, my presentation touches on extensions of its use with other kinds of documents. It has been written with the intention of being open-ended and flexible enough to accommodate the addition of routines tailored to address needs not met in its current configuration.

IVANHOE: Humanities Education in a New Key

Jerome McGann, The John Stewart Bryan University Professor, University of Virginia

Introduction.

IVANHOE is a research and pedagogical project for humanities scholars and students working in a digital age like our own, where books are only one among many cultural sources and objects of critical reflection. It is designed within the framework of the traditional goals of humanities education: to promote rigorous as well as imaginative thinking; to develop habits of thoroughness and flexibility when we investigate our cultural inheritance and try to exploit its sources and resources; and to expose and promote the collaborative dynamics of all humane studies, which by their nature both feed upon and resupply our cultural legacy.

IVANHOE emerged in the spring of 2000 from a conversation between Jerome McGann and Johanna Drucker on the subject of literary-critical method, and their shared dissatisfaction with the limitations of received interpretative procedures. They were interested in exploring forms of critical inquiry that moved closer to the provocative freedom of original works of poetry and literature.

McGann suggested that Walter Scott's famous romance fiction *Ivanhoe* contained within itself many alternative narrative possibilities, and he added that this kind of thing was characteristic of imaginative works in general. Scott's book epitomizes this situation in the many continuations it spawned throughout the nineteenth-century – versions in different genres as well as other kinds of responses, textual, pictorial, musical. For example, when many Victorian readers complained about Scott's decision to marry Ivanhoe to Rowena and not Rebecca, they were clearly responding to one of the book's underdeveloped possibilities. In our own day readers often react to other unresolved tensions in the book – for example, to the complex ways it handles, and mishandles, the subject of anti-Semitism.

"Everyone knows that an anti-Semitic strain runs through the novel," he said. "The question is: 'What are you prepared to DO about it. Victorians rewrote and reimagined the book. Why are we so hesitant about doing the same thing?'"

The concept of criticism as "a doing", as action and intervention, is a founding principle of IVANHOE. Traditional interpretation is itself best understood as a set of reflective activities and hence as something that lays itself open to active responses from others. It is not so much that "all interpretation is misinterpretation", as Harold Bloom observed some 25 years ago, as that all interpretation pursues transformations of meaning within a dynamic space of inherited and ongoing acts of interpretation. Interpretation is a dialogical exchange and, ultimately, a continuous set of collaborative activities.

This critical vantage necessarily resists the traditional assumption about the self-identity of a particular text or cultural work. Various factors and agencies so impinge on the textual condition that the field of textuality, including all the objects we locate in that field, are in a perpetually dynamic state of formation and transformation. This view of textuality implies that any textual object -- what in IVANHOE we call "the source text" -- has to be encountered within a dynamical "discourse field" (i.e., the extended network of documents, materials, discussions, and evidence within which the work is continually being constituted). Approaching textuality in this way, we concluded that a digital environment would provide IVANHOE with an opportune and useful playspace.

When we began playing IVANHOE these initial premises were a somewhat loosely held set of intuitions. The actual gameplay transformed them into clear and governing ideas. Not surprisingly, it also drove us to rethink the whole process of interpretive method and theory. As a result, we began to see that IVANHOE could be designed and developed as an environment for the study and encouragement of critical practices that would make self-awareness pivotal to the whole enterprise. IVANHOE is what Coleridge might have called "an aid to reflection": a machinery for making explicit the assumptions about critical practice, textual interpretation, and reading that remain unacknowledged, or irregularly explored, in conventional approach to literary studies.

In IVANHOE, the idea is that interpretation should no longer be imagined as proceeding from a subject grappling with a transparent object. By contrast, IVANHOE discourages players from assuming that there is something to be called (say) "The Poem Itself". Perhaps even more crucially, it routes the acts of an interpreting agent back into the material being studied. Players and their moves are continually returned to the ongoing process of collaborative investigation for further critical reflection, both by the agent herself and by the others players. All players thus move in that Burnsian space where each is repeatedly drawn "to see ourselves as others see us". Based on economies of expenditure, deficit, and gain, with winning conditions and costs, IVANHOE's underlying game model urges the player – the *thinker* – to a continuing process of measuring and assessing his or her moves in relation to everyone else's.

IVANHOE has been dominated from the start by a ludic spirit. This attitude is reflected in the name of the project, IVANHOE, which references a cultural work now rarely taken "seriously", though it was once reigned as perhaps the most popular and seriously influential work of fiction in nineteenth-century Europe and America. We took that avoidance as a sign of a poverty of criticism, which goes broke by following a Gold Standard of value. IVANHOE would encourage, instead, as much circulation and exchange as possible.

From the initial provocation, IVANHOE quickly spun itself into life. Playing with Scott's novel generated new practical design features, the most important of which was the idea that the game would have to be played "in" a role, or *en masque*, under an explicitly assumed conceit of identity. Players would make their moves only through that role. This device would introduce into IVANHOE another vehicle – in addition to the dialogic form and performative procedures-- for encouraging critical self-reflection. We also began to see that a robust environment would only be built if we tested our ideas in as many kinds of gameplay as possible.

The Development Schedule.

IVANHOE was initially conceived as a general purpose tool for enhancing a person's range and acuity of critical reflection on some given set of cultural material. The first test iterations of its use focused on particular, paper-based literary works Scott's *Ivanhoe, Wuthering Heights*, *Frankenstein*, several stories by Murakami, "The Fall of the House of Usher", "The Turn of the Screw", and Blake's *The Four Zoas*. These sessions were run to test IVANHOE's performative methodology and its logical structure, and to clarify the requirements needed for building IVANHOE as a digital environment.

Played in what was essentially an electronically enhanced paper-space, these iterations were most successful in the ways they exposed the critical and interpretive power of performance-based acts of textual invention. They supplied us with useful information about how to construct an initial IVANHOE design for studying traditional text-based materials. The test runs also suggested two other useful ways in which to explore the tool's design possibilities: first, to deploy IVANHOE as both a pedagogical and a scholarly research tool; second, to launch its functions in a born-digital database of materials. IVANHOE's interpretational capacities were conceived to have wide range and flexibility across every sort of informational material in the humanities and the social sciences

In the past year (2003-2004) the online playspace has been in development and we have run test sessions with Poe's "The Fall of the House of Usher", Hawthorne's "Rappaccini's Daughter", D. G. Rossetti's "The Orchard Pit", Swinburne's "A Criminal Case", and Jerome K. Jerome's *Three Men in a Boat*. These sessions were run to clarify the technical and interface issues and lead us to implementation. The 1.0 release of online IVANHOE is scheduled for 1 November 2004.

"What Does a Session of IVANHOE Look Like?"

A group of people, two at a minimum, agree to collaborate in thinking about how to reimagine a particular work, say *Ivanhoe*. The agreement is that each person will try to reshape the given work so that it is understood or seen in a new way. The reshaping process in IVANHOE is immediate, practical, and performative. That's to say, the interpreters intervene in the textual field and alter the document(s) by adding, reordering, or deleting text. Interpreters are expected to produce a set of interventions that expose meaningful features of the textual field that were unapparent in its original documentary state. Interpreters will also look for ways that their interventions might use or fold in with the interpretive moves of others working the collaborative session of IVANHOE.

An IVANHOE session typically extends for a set period of time – we have found that a week seems a useful timeframe for pedagogical purposes. To this point we have played IVANHOE by focusing on single literary works. Nonetheless, a session of IVANHOE might focus on a set of works that define an interesting cultural phenomenon – The Salem Witch Trials, for example – and in that case the pedagogical event might well run on for a much longer period.

Some analogies may be helpful for understanding what it means to play IVANHOE. Interpreting agents in IVANHOE approach their work much as performers or conductors approach a piece of music, or the way a director approaches a play. Performance fashions an interpretation of the original work, and the result is what Gertrude Stein, in a slightly different sense, called "Composition as Explanation". Performative interpretations of all kinds – translation, for example – have much in common with IVANHOE. Book artists and illustrators work along similar interpretive lines, and we have many cases where authors themselves illustrate or design the embodiments of their own textual works, thereby glossing them with intervening sets of interpretive signs. Some notable figures integrate text and visualization into a composite or double work – in England one thinks immediately of Blake, Dante Gabriel Rossetti, Edward Lear, Lewis Carroll. Or consider how "The Matter of Arthur" or "The Matter of Troy" are conceived and elaborated. A set of legends centering in the Trojan war and in King Arthur multiply as versions and variants that expose fresh ranges of meaning resting latently in the materials. The interpretive transformations that unfold in a session of IVANHOE seek to exploit a logic of interpretation of those kinds.

IVANHOE is not like a "creative writing workshop", however. Its textual transformations get executed in a frame of reference focused on the significance of the changes *in relation to the originary textual field and the changes that one's collaborating agents make to that field*. The presence of the initial state of the text(s) is always preserved because the point of IVANHOE is to study that field of relations as it provokes or licenses its readers to reimagine its implications and textual possibilities. Interpreters are expected to keep a journal in which their interpretive moves are justified and explained in relation to the originary work and/or the moves made by the other agents.

IVANHOE and Scholarly Research.

Research Example 1. IVANHOE is built to function as a collaborative workspace for scholarly research in some specified topic area. Its dynamic interface is especially suitable for integrating research work that utilizes online scholarly resources -- for example, the large database of materials in *The Rossetti Archive* (housed on the IATH server at University of Virginia). IVANHOE can organize a joint scholarly workspace uniquely suited to exploit the research potential of databases like *The Rossetti Archive.* In the case of the latter, for instance, a group of scholars might undertake to study the relation of aestheticism and modernism through the crucial (and neglected) case of Rossetti and his circle. Drawing on their expertise in art history, literary scholarship, bibliocriticism, cultural studies, and graphic arts, scholars would address questions like "How does the aesthetic of Rossetti's work challenge received concepts of visual and literary modernism?"

The first round of moves for such a gameplay might require each player in turn to identify, and link for study, a series of documents housed in *The Rossetti Archive*. These nets of material would represent the point of departure for each player's subsequent critical moves, which would make connections with other documents within and without *The Rossetti Archive*. (An apparatus for maximizing that critical process has been invented by Bethany Nowviskie and is currently under development as well.)

As in any version of IVANHOE, each participant assumes a role. In this specific case, the role is the clearly defined professional perspective each brings to the project. Thus, in her player file the graphic design historian glosses her moves with commentaries relevant to this particular critical approach. These glosses represent her critical reflections on the game play and especially on her own moves in the game. The glosses constitute a rationale for the relevance of graphic design to Rossetti's work, to Pre-Raphaelitism, and to their place within the culture of modernism.

Moves are made by adding links, files, and new materials in order to develop and explore lines of critical argument. Ultimately, the form/format of the final project will emerge as a visual and textual network of documents, files, and supporting materials that comprise a response to the original research question. The final work will read as a series of screens with links, commentaries, and annotations by individual players. If the research project is conducted in a pedagogical context – for instance, as part of graduate course work – a premium point system would operate for moves that draw critical relations (in the form of links) with gameplay being pursued in areas outside the player's role expertise. In any case, a synthetic rationale should pervade this kind of deployment of IVANHOE. In exploring any particular disciplinary area, players will be seeking ways to connect their critical networks of material with the networks being developed by the other scholars. These connections might be integral or differential in character, drawing patterns of relation via perceived similarities or perceived distinctions.

Each scholar's player file will thus build a meta-commentary that relates the specific scholar's work in IVANHOE to the work of the others. Readers coming to the work can then follow the overall argument, the development of the discussion through the logged/threaded discussion, or follow any of the particular player's individual discussion in their player file.

Working from its awareness of the logical structure of *The Rossetti Archive*, the computer will output, at the end of the gameplay, a representation of the critical patterns it is able to distinguish within the entire set of linked materials developed by the scholars. The project concludes with a real-time seminar in which the scholars make their own assessment of the outcome of the exercise. The entirety of these materials – the moves, the player files, the computerized representation, and the concluding scholarly assessment – are, finally, uploaded into *The Rossetti Archive* for future scholarly uses.

Using IVANHOE in this way, scholars will be able to work collaboratively without having to subsume individual differences of interpretation or opinion. New materials pertinent to specific areas of expertise will be uploaded and linked by individual players to support the interpretive network each player is trying to develop. Though finite and bounded as an interpretive constellation, the final work of each player will thread through and interweave with the other players' interpretive nets. The experiment demonstrates the interplay of collaboration and individual interpretation in scholarly activity, as well as the functional capacity of IVANHOE to provide effective access to an electronic archive for scholarly research purposes.

Research Example 2. An advanced graduate seminar will use IVANHOE in the spring term of 2005 at UVA to undertake a scholarly research project in two large databases, *The Rossetti Archive* and *The Blake Archive*. The question to be addressed will be: "How do the works of Blake and Rossetti converge with or break from Romantic tradition?"

The seminar will be divided into two groups. Group One will play IVANHOE in *The Blake Archive*, Group Two in *The Rossetti Archive*. Each group operates, in this respect, as a collaborative research group (as described in the previous example of using IVANHOE as a collective research environment). In this case, however, each group will also "lurk" in the game play of the other group. As lurkers, each group will have the responsibility of assessing the game play of the players in the other group and of that group as a whole. The assessment will come as a final report on the other group's game play.

IVANHOE and Pedagogy

In recent years the scene of humanities instruction grows less like the classroom of the 1930s, when the remarkably successful teaching protocols of the New Criticism were invented. New Critical pedagogy centered in a single textual object – "The Rhyme of the Ancient Mariner", *The Symposium*, *Pride and Prejudice* – that would be brought into the classroom for close reading and discussion. That model for a classroom procedure was so effective that it still dominates the way the humanities classroom is conducted in high schools, colleges, and universities. Indeed, its procedures remain in certain ways foundational to any kind of effective education. But our classrooms now are populated by students for whom the book is only one kind of communication tool. Like ourselves, they live every day in a complex communication network of paper as well as electronic texts, and of texts as well as all sorts of other media, much of it mixed. Because we all bring that world with us into the classroom as (so to speak) the cultural air we breathe, New Critical models of instruction now regularly specialize and restrict both the materials and the arena of that general education the humanities educator has always so carefully cherished. Because the humanities has never been about specialization but about the training and education of broadly informed citizens, we are being called to imagine new instructional methods and procedures. IVANHOE is being developed to help answer that call.

Research in the field of education has made a convincing case for the use of games in promoting goal-centered learning. Ivanhoe makes use of selective principles of role-playing scenarios – such as requiring players to chose a real or fictional identity and create their interpretation or analysis from that point of view. Likewise, role-playing is an established practice in the constructivist classroom. Ivanhoe makes use some of these features of entertainment and game models to motivate reading, interpretation, and study of documents that are traditionally associated with the humanities. Most fundamentally, Ivanhoe seeks to promote self-conscious critical thinking.

Ivanhoe works by encouraging players to work with a designated textual work and its sources, variations, versions, and other materials relevant to the history and production of the text. At a basic level, this will encourage such activity as the comparison of an illustrated version of a classic work to a text-only edition, or a facsimile manuscript and a printed edition. Students will be introduced to the concepts of bibliographical studies, theoretical issues in textual interpretation without having to first engage with a technical vocabulary. Ivanhoe allows them to enact the principles of comparison and critical analysis that are essential to the humanities and social sciences where informed qualitative judgments are crucial. Collaborative, peer-exchange models of engagement will encourage cooperative development of analytic skills in reading and comprehension and appreciation of individual points of view in writing. Ivanhoe promotes curricular dependence on creative, synthetic practices and engagement with primary materials that have traditionally been inaccessible in classrooms.

In a post-secondary context, players will be encouraged to develop library and research skills through the integration of traditional text-based materials and on-line resources for playing the game. A great degree of self-conscious awareness, and a higher level of bibliographical skill will be required. Crucial skills in assessing the validity and credibility of sources and self-conscious awareness of the point of view from which a player makes a critical judgment will be encouraged by the structure of the game. Players will rewarded to the degree that their critical interpretations have been made explicit within an interactive community of other players by creation of well-documented commentary on their individual contributions, and critical assessment of other players' work.

Conclusion.

In summary then, IVANHOE can be used in a variety of ways as a competitive, game-like environment, as a collaborative study and research situation, or as a context in which players strive to achieve their own individual goals. In a classroom setting, IVANHOE could encourage students to improve bibliographical and research skills in one round and critical reading skills in the next. Individual students could decide which of several interpretive skills they wish to improve in a round of play, or they could consult with a teacher to set these goals. For more mature players, various competitive or collaborative situations might be imagined to promote specific types of critical reflection and scholarly research.

IVANHOE can be played in a game mode with points, scoring, and competitive interactions. It can also be used for non-competitive collaborative work within a community of scholars or in classroom activities.

It is important to note that although developed from models taken from literary studies, IVANHOE is not subject specific, and can be readily adapted to the questions that are a regular concern throughout the humanities and social sciences. Rather than operate as a delivery mode of pre-packaged content, it is a tool that can be configured anew by instructors and scholars according to the goals that suit their research or pedagogical circumstances. It is an effective web environment for any field of cultural investigation that is primarily document and text-based, and in which access to electronic archives, collaborative work, and critical interpretation are central concerns.

Modelling computer assisted reading and analysing of texts (CARAT) : the SATIM approach.

Jean Guy Meunier, Université du Québec a Montréal
Ismail Biskri, Université du Québec à Trois-Rivières
Dominic Forest. Université du Québec a Montréal

Introduction : the problem

Since now practically fifty years technologies for computer assisted reading and analysis of text (CARAT), have penetrated the various humanities and social sciences disciplines. Since now practically fifty years technologies for computer assisted reading and analysis of text (CARAT), have penetrated the various humanities and social sciences disciplines. One finds them in philosophy (Mc Kinnon, 1979; Meunier, 1976; Lusignan, 1985), in psychology and sociology, (Barry, 1998; Alexa et Zuell, 1999a, 1999b; Glaser et Strauss, 1967), in literature (Brunet, 1996; Hockey, 2000; Fortier, 2002, Bradley and Rockwell 1992. Sinclair 2002) in textual semiotics (Ryan, 1999, Rastier, 2002), social sciences (Jenny, 1986, Fielding and Lee 1991, Lebart et Salem, 1994, Mayaffre, 2000), in history (Greenstein , D. ,2002) ,etc.

This technology is different from the AI approach to discourse analysis (Hobbs, 1990, Marcu, 1999, 2000) or automatic reading (Ram and Moore 1999) where the objective is to have the computer simulate some type or other of "understanding" of a text in some specific application or process (inference, summary, knowledge extraction, question answering, mail routing, etc.) . It is also different from information retrieval (Salton and all 1995a, 1995b) or hypertext technologies (Rada, 1991) where the objective is finding documents for a particular query or navigating through documents. In CARAT, literary critics, philologist, content analyser, philosophers, theologians, psychologists, historians etc. and even any other type of professional text readers (lawyer, journalist, etc.) will not accept automatic text "interpretation" tools under any form whatsoever. They require instead sophisticated tools to assist them in their own process of interpretation be it reading or analysing.

And computer technology has been offering more and more possibilities to assist these reading and analysis processes. To assist the reading process, the computer technology has been offering more and more possibilities. Archive of electronic texts are now a common thing: V. g. : *Oxford Text Archive*, the *Brown Corpus*, *Perseus*, *Gallica*, *Frantext*, etc.) They are more and more critically edited, standardize, (SMGL, XML HTML, etc.). and can be explored with "intelligent" tools such as navigators, research motors, hypertexts etc.

To assist the analysis process, the technology has also been very constructive. In fact, one could distinguish various generations of this technology. A first one (1950-1970), opened the era of the capture of the electronic text. A second generation (1970-1980) offered tools for the description and manipulation (extraction, concordance, statistics, lemmatization, etc) of these electronic texts. A third one, (1980 -1995) started a standardized tagging (TEI initiative) and the linguistic processing (syntax, semantic, discursive and rhetorical, etc.) of the text. A forth one (since 1995), introduced sophisticated mathematical models (classifiers, neural nets, categorizers, etc.) on the text. There exist today a proliferation of such tools: from the more hybrid to the very specialized, from the laboratory prototypes to the more industrial applications. (vg Concord, Word Cruncher, Tact, Cobuilt, Word Smith Tools, Lexa, Cosmas, Tropes, In-text, Alceste, Sphinx, Hyperpro, Atlas, Nu*DIST) CARAT has even received a buzz name in the commercial field: Text Mining, for which there exist a multitude of tools.

Unfortunately, the traditional communities of the humanities and social sciences, except maybe for critical editions of texts, qualitative content projects or historical archiving, do not recognize the importance of this technology. In fact, it has not really been integrated it in its practice. In the majority of cases, reading and analysis is still done in a very classical manner. Many researchers have tried to understand the reasons of this attitude. In 1992, S. M. Hockey noticed that these technologies have not obtained the same success in the humanities and social science as in other sciences. One reason advanced is that they were not really adapted to the dynamics of reading and analysis of text as practiced by them. In fact, even though the computer technologies are accepted more and more for classical, transcription, navigation, and configuration of texts, they encounter many difficulties when applied to CARAT. Many researchers believe in it. However, they are still unsatisfied with the methodologies and tools available.

One can invocate many reasons to explain these difficulties. A first one is the weakness of the ergonomics of these technologies that renders their learning and usage difficult for many users, particularly in the humanities and social sciences (Unsworth, 2003: 2). Except for a few systems, many technologies have often developed by communities external to CARAT (linguistics, artificial intelligence, data processing, information retrieval, etc.) and for specific objectives. A second one relates to the limited sets of tools available for text analysis, where often , a researchers is confined to text transcription ,text encoding, lexical (concordances, collocation) basic linguistitics (lemmatization, tokenzer, morphological taggers) or statistical analysis. Thirdly these many commercial tools are often geared more to information retrieval than real assistance for reading and analysis of text. Finally, the technology is often a closed one. It is also proprietary, rigid and non-modular and, as often underlined, not really adapted to the dynamics of projects in CARAT.

We believe though that there are more profound theoretical reasons for these difficulties : 4) there is a lack of understanding of the effective role played by this technology. The mean stream community still entertains the image that it is the computer that does the text interpretation instead of seeing it as an assistant in this process. 5) There is also a misunderstanding of the linguistic aspect of what is a text, its reading and analysis. Rastier (2000) in one that readily reminds us how much a text as a whole is highly different from a sentence of a language and must be approached with tools different from grammar and logic. (Marcu, 2000. Mann and Thomson, 1988) 6) Finally, the classical technologies of CARAT, except for the ones of standards and protocols (Sinclair 2002) have not really integrated the modalization tools of the software engineering languages as developed in the last 15 years. (Braumbrough, 1991; Levesque 1998) and which would allow modularity, reutilisation, exchanges, collaboration and even automatic code generation.

So, if a renewal of this technology is to happen then, one must have a more rigorous idea of what is reading and analyzing a "text" with the help of a computer. That is: a more theoretical and formal foundation of CARAT technology has to be explored.

The various generations of the technology have offered computer design for CARAT. But they were often ad hoc solutions, resolving specific and limited problems, delivering heuristic applications in various fields but lacking in integration. So, if the technology requires a renewal, it has to be founded on a more complex and integrative design. And, because of the complexity of such a problem, we must explore more subtle design of CARAT. In this paper, we shall briefly expose the SATIM approach to this problem.

The design proposed here is presented in three levels : a first one, the *analytic*, describes CARAT process in terms of the methodologies used in various social and humanities practices ; a second design analyses this same process in a functional language; and finally, a third one sees it in terms of the procedures that can be realized in a computer architecture.

The merging of these three designs, we hope, allows a more integrated understanding of the CARAT process. The three designs see it as a cognitive path of an interpretative or semiotic nature which an expert reader will follow when processing textual data in order to produce knowledge; they place the computer in the secondary position of a technological assistant in this process.

The analytic modelling

The first level of modelling design, is a analytical one. Its aim is to identify, in a descriptive but semi formal language the constituents of a reading and analysis process as practiced by the experts. Indeed, each expert of one or other discipline applies in his reading and analysis of text a methodology that has been acquired in various applications . Each discipline has distilled from its own practices a certain number of processes that have been adapted for computer processing. If a more refine computer architecture is to be built to assist one or other methodology, an analytical description is required so as to identify the basic and pertinent components. In this analytical design, one will have to identify three main components: units of information, operations and processing chains.

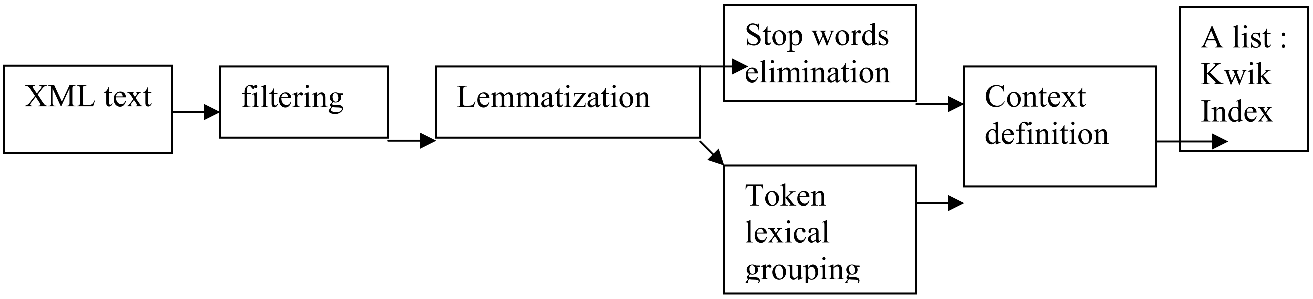
The first components of the analytical model are the **units of information**, that one has to identify the basic if not the atomic components of the processes. Traditional approaches usually take the word as the basic unit. But, as it has been repeated (Benveniste 1966) this is a highly laden theoretical choice. For what is a word? And in CARAT, as we shall see later on, there can exist many types of basic units of processing. That is, units of information (UNIF) may be more or less granular depending of the methodology used. They can be also parts of words, whole words , sentences , paragraphs etc . They are in fact are the input and output of various operations and processing chains.

The second components of any reading and analysis process are the operations applied to the these units of information are the tasks that each expert sees as basic to its methodology. Although they often appear simple they are in fact very complex interpretative operations, often composed themselves of many sub-operations. However, the experts accept them as components or steps in the methodology even though they may not be aware of all the cognitive and algorithmic details underlying each operation.

Here are examples, of traditional CARAT operations applied on units of information such as words and sentences: : *The cats are on the mat.* .On this sentence, one can apply a lexical extraction , a morphological extraction , a lemmatization , a counting, various types of tagging , etc. Here, each operation has a specific type of input and delivers a specific output: Some operate on words, others on phrases, paragraph, and even whole texts. Here a few more of these operations:

As it can be seen, the input of each operation can be various linguistic semiotic forms of the original text (words, n-gram, sentences, paragraphs, chapters, fragments, documents, etc.). Each operation is not a simple task; it may include many sub-operations. Some of them can be only realized manually others can be assisted by computer. There exist a myriad of such operations. And, any serious analytical model of CARAT will require the identification of these basic operations for each methodology. They constitute the building blocs of a solid CARAT analysis chain.

The third component of A CARAT application is the processing chains. Indeed a X CARAT analysis is not simply a set of such operations. It is also a combination of such operations. Each of which can form what we a call an *analysis chain*. In our technical terms, a analysis chain is a combination of operations but where each output of one operation can be the input of one or many other operations as in the example 3 and 6 where the output of the lemmatization operation is the input of the XML tagging. Basic operations can by this composition become more complex operations. So that, in fact, a particular CARAT methodology, becomes in turn some particular combination of operations. For example, a thematic analysis is a complex combination of many sequential and parallel operations on a text. Here is a simple example of such an analysis chain: A concordance chain:



In this example, we have chosen but a few operations. Many others are possible depending on the objectives aimed at. And finding good and pertinent combinations is often at the core of a research project. Even if there exist a stable methodology and some good practices (cf qualitative, terminological, stylistic, categorization, discourse analysis, etc) many combinations may have to be tested and validated before one is accepted. How for instance does one build such a chain for authorship identification? What is the best thematic navigation course in a text? It is precisely here that granularity, modularity and flexibility in the definitions of the operations are important. Combinations depend on their quality.

By making clear the systematic nature of the operations, we are in a better position to determine which series of operations and especially which sub-operations can be carried out by a the experts and which ones can be given an algorithmic translation. The analytic modelling becomes a founding stone for a better CARAT technology. However, before identifying some of these dominant operations and analysis chains, we must give these concepts of unifs, *operations* and *analysis chains* a more formal definition.

The functional Design.

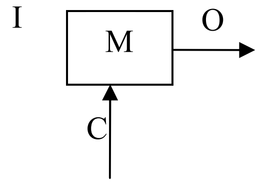
In the analytic model, one aims at defining the set of operations and analysis chains explicit the components of a CARAT methodology. In the functional modelling, these operations and analysis chains are now translated into functional terms. This translation allows a more formal definition of an operation in terms of their internal structure and their relations with others operations. Finally, this translation opens the ways to a better specification of the third level of design: the computer architecture.

In the analytical model the basic components were the *units of information*. These units are their inputs but also the outputs of the operation. Hence, they define a domain and a field on which the operations are applied. We can distinguish types of such units . For instance, *words* are an object -type different from a *sentence*-type. For instance, they may be words defined as sequences of characters separated by blanks. But they may also be a lemmatized, words (book (s)), morphological tagged word (book as name vs Book as verb) complex words (vg : book keeping) or they may be n-gram as sequences of characters (vg. books: Boo-ook oks ; (Damashek, 1989). In higher lever of analysis they may be wholes clauses [Grimes, 1975, Longacre, 1983], prosodic units [Hirschberg and Litman, 1987], turns of talk [Sacks et al., 1974], sentences

[Polanyi, 1988], intentionally defined discourse segments [Grosz and Sidner, 1986], etc. Distinguishing various types of objects allows the building of classes of operations that have same type of input and output objects. This in turn allows a higher level of generality for the CARAT functional modelling.

The second component of the analytical model were the operations: In the the functional model, these operation will now be translated into formal functional structures. In the mathematical sense, an operation has an internal structure. This means that an operation is in fact a function that relates objects of some type to other objects of another type. However, this relation is realized in certain way. The function must be effectively applied (through a procedure) that is constrained in some manner or other. It follows then that a function is described by its input (domain) and its output objects (field), its procedure, and controls parameters. The domains and fields of a function can be of various types. They can be objects (vg numbers, linguistic symbols, etc) or functions themselves. The procedures can be non-computable or computable operations applied to the objects (vg move, go to, add, delete, etc). The controls are the conditions under which the procedure is applied. (v. g. the rules and parameters of a procedure).

We can represent the structure of a function by the following graph:

	Where : -I is an input: the *domain* of objects on which the function is applied (its arguments). -M is a procedure: effective *operations* that are applied to the objects. -C is a control: the *constraints*, rules or parameters of the operation: -O the output: the *field* of the function or value of the function: (the value of a function).

In CARAT, the objects of a function are the UNIFs. But, they can also be some other functions. Unfortunately, in many applications, the procedures and controls (rules or parameters) are often not always explicit (except in the program itself). For example : a grammatical rule for parsing a sentence will mix the procedure and the parameters of its application. Here are simple examples of some CARAT functions:

1- A *lexical extraction* function is a procedure whose domain (I) is a set of words (sentences, paragraphs, etc), and whose operation (M) consist in extracting according to some rules (C) and the value (O) of the function is a list of the word types. V. g. : Input : The cat is on the mat ==> Output: { The, cat, is, on, mat}

2-A *tagging function* is an procedure that takes two sets: a set a linguistic sign (I_1) (vg "Peter") and a set of tags (I_2) and delivers (according to some rules C) a doublet ((Peter), (Proper Name)).

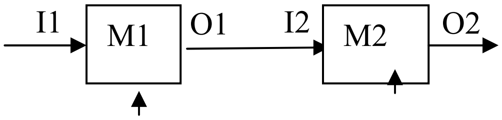
3- A *lexical counting* function is a procedure that takes two sets : a set of words in a text (I_1) and a set of numerals (I_2) and according to a cardinality count delivers a doublet that gives for each type of word the number of its occurrence vg. . The tagging operations become then becomes a tagging function. V. g. : Input : The cat is on the mat ==> Output: { (The, 2), (cat, 1), (is, 1), (on, 1), (mat, 1)}.

The translation of operations into functions is not but a change of name. It allows rendering more explicit the internal structure of the operations identified in the first level design. That is, the domain of which there are applied, their procedure, their control and their results. From then on, it becomes possible to compare the functions among themselves, either from their input, their output, or their procedures and controls. Some functions are then seen identical because of they are applied to a same type of objects, others because they present a same type of procedure or control.

For example : an *lexical* counting operation becomes similar to a *lemma* counting operation if one accept that there words are from this point of view a similar type of objects as a lemma. In this sense, many operations can be similar because they operate on the type of objects: words, n-grams etc. But they may be distinguished on the grounds of their rules or control parameters: For instance a syntactic tagging of a word is an operation similar to a semantic tagging. Both take as input words and add a particular tag to it. They differ by their constraints on the operation that affects their interpretation.

Analysis chains as composition of functions.

The third components are the analysis chains. The translation of operations into a functional language allows us then a more formal definition of the combination of functions giving rules for their compositionality: that is their well-formedness. Indeed, each function can either take as input, the output of another function. This allows building complex functions. These functions can then be combined with other functions: by composition of functions.



But, because each function it is in itself an autonomous object and is well typed, it can become the input for other second order functions. This allows a categorical structure of compositions such as found in the categorial grammars (Shaumyan 2002 Desclées, 1990, Fielding and al. 1991, Biskri and Desclées 1997) many sorted algebra (Montague 19741972) of combinatorial logic (Church 1941, ; Curry 1958, and Feys 1974). In the SATIM approach (Biskri & Meunier 2002) we are exploring the possibilities to formalize the composition of functions of an analysis chain in terms of operators as defined in combinatory logic. This allows a perfect control of structure of the chains . Analysis chain then becomes syntactically well formed chains of complex functions. For example , in the preceding example, we find a sequence a two operations . in a combinatorial grammar, , through specific rules of combination (combinators) the whole sequence can be itself be transformed into a new a well formed higher level complex function. For instance, we have now a new complex function composed of two functions: M1 and M2 that whose input are of type a and deliver in syntactic manner a result of type c. So, in a serial chaining of two functions, we can, by these composition rules we can build systematically complex functions out of more elementary or granular functions. The new complex function will be itself considered as an elementary function for another chain. In this case, a patrimony of existing functions can be recuperated, saving hence time and energy.

Some functions in CARAT.

We shall now present some main functions in CARAT. These are inscription, classification, categorization, exploration.

Inscription.

The first type of functions are the ones that relate a document outside of CARAT so that is can be processes into the CARAT chain. Normally, a text comes as a document whose physical carrier can be paper. But it could also come in a electronic form (Internet HTML, PDF, Word processor) etc. All text reading and analysis will require that these external documents be transformed, into a format accessible for CARAT processing. Some of the typical operations of these inscription functions can consist of locating and manipulating physical documents (manuscripts, monographs, index cards, etc.) so as to compose the corpus of textual information that will be submitted to a reading and analysis. For instance, if it comes in a traditional paper, photographic or electronic image it may require many complex sub-operations such as *scanning*, *OCR processing*, and afterward *editing*, *cleaning* (vg, non pertinent information), *correcting*, etc. *ASCI translations*, etc. Even if it comes in an electronic format, produced be various text processors or other electronic text manipulators it will often required complex sub-operations such as *filtering, XML tagging, restructuring*, etc.. In other words, one needs a set of inscription functions that transforms textual document into a form that is admissible to CARAT processing.

Classification functions.

Once a document text as become admissible to some type or other of CARAT processing, it will be manipulated in different manners, all of which can be seen under one main type: a classifying function of some sort or other. Some are very simple other more complex. In general terms, a classifier is an abstract machine or function that realises some type of grouping or sorting of objects. In a set theoretical definition, classification is the projection of a partition on objects. So that, they are as homogenous as possible. In functional terms, a classification is an equivalence function applied to the objects: They must fall under one same predicate. What differentiates one classifying function from one other are the control parameters or rules. This may not be very apparent at first sight but a formal analysis of many functions will reveal that they are in fact some type or other of classification. For instance, the addition operation may not look like a classification, but if translated into abstract algebra, the ADDITION is a functional relation between a set of doublets on numbers into a set of numbers. In other words, the ADDITION operations forms a special type of class called a ring with specific properties.

CARAT offers also such classification operations. It may not be apparent at first sight that many functions are classification. In fact, many of them are grouping, clustering, listing operations on a set of symbols, Here are examples of such algorithmic classification functions.

-*Lemmatization*: builds set of words with some common part defined by a rule.

-*Lexicon* extractor: It put the token words into their type and produces a list as a set of word types.

-*Concordance*: build a set of fragments where one key words has it center.

-*Clustering word*: builds set of textual fragments that share some common structure depending of the similarity criterion (same words, etc).

-*Complex words Identifier*: by some rule forms sequences of word (as an ordered set). (eg. object-oriented-relational-data-base, real-estate-agent, etc.)

- *Sorting documents :* builds class of documents by some similarity measure.

Some of these functions may be computable and hence may receive algorithmic translations. But some may not be computable, either for theoretical reasons or for complexity reasons. They will hence have to be done manually. This does not alter their functional status!

Categorization.

Highly related to a classification function are categorization functions which are very important in CARAT. A categorization function is a classification function that is applied to sets of objects of different types. In a classification function, the operation groups the objects into a class. In a categorization function, the function also groups the objects but related them to element of another type or set of objects. In that sense, a categorization is a special type of classification.

In CARAT, we can find myriads of such operations.

- Edetic: XML, SGML are categorization operations between parts of text (word, fragments, etc) and a set of categories.
  - Syntactical: associates to word their grammatical function,
  - Semantic: associates to word semantic descriptions, etc.
  - Pragmatic: associates to words or sentences their users, addresses, etc.,
  - Argumentative: associates to words or sentences their presuppositions, arguments, etc.,
  - Logical: associates to sentences their logical role: premises, minors, conclusion, etc.,
  - Hypertextual: identifies relationships with other parts of texts, (Rada, 1992; Conklin, 1987).

- Documentary: associates to words or sentences their indexical role, etc.,

- Interpretative: associates to words or sentences a sociological, psychological, anthropological category etc.)

-Co-occurrence : associates to words or sentences their immediate lexical environment.

-Counting : associates to words or sentences set a cardinality.

Each reader and analyst will obviously have his own categorical framework. A literary expert will associate categories that are different from those of a sociologist, a linguist, a philosopher, etc.

The operation of categorical description may be carried out in several ways. In order to understand this operation, one must distinguish it from a certain number of other operations.

First of all, categorical description must be distinguished from annotation. Annotation (Nielsen, 1986, Marcu, 1998) consists basically of a direct addition of information to the text, inscribing various types of information on the texts. (comments, relationship, corrections, references, labels, etc.) in the form of notes. Annotation often produces an independent text. Contrary to annotation, categorization adds, to the units of the text, clarifications concerning their functions, in accordance with an interpretative template. Categorization projects a specific structure onto the text. It "marks" it with specific structural information and attaches to more or less complex labels identifying some of their properties.

Secondly, categorization must be distinguished from tagging. Indeed, tagging is but one among many sorts of possible categorization. Its real aim is to project on set of objects a structure of some sort, that is, to relate set of objects with other set. The tagging is but the naming of part of this structure. For instance, one could structure the following sentence:

John loves Mary into *((JOHN) (LOVES (MARY))*

This would render explicit the categorical structure of the sentence. But tagging it which could make even more evident such as:

$(\text{John})_{\text{name}}$ $((\text{Loves})_{\text{ Verb}}$ $(\text{Mary})_{\text{ name)SV}})_{\text{S}}$

Finally, a third distinction must be made between the categorization function and the computing strategies which carry it out. Categorizing is a type of classification.

The two preceding type of categorization, structural and simple tagging, are not on the same level as to their computationality. In fact, mathematical theory has demonstrated that not all forms of classification are computationally definable; there is not always a recursive function which defines them. For example, a categorization which associates socio-political labels (eg. government, institution, deputy, and elector) with expressions, will not be part of a grammar, whereas syntactical categories can belong to grammar. Even in cases where classes can be recursively enumerated, there may not be any algorithm to express them. As far as texts are concerned, this means that there does not exist necessarily recursive functions and algorithms for all the categorization operations, or if they do exist, they are not yet necessarily known.

Because of these distinctions, categorization must be seen as a complex operation which, in some cases, cannot always be carried out by automatons. This means that the reader and analyst must often do it manually or semi-manually, occurrence by occurrence. For example, only one reader can say can that in a particular text the expression *Bonaparte* belongs to the category (CHAUVINISTIC EMPEROR). The choice of categorization strategies and categories then depends on the reader's and analyst's intentions: a terminological analysis is not a stylistic analysis, which in turn is not the same thing as a study of argumentation or an indexing aid.

Configuration functions.

The last set of important function in CARAT pertains to the cognitive relation between a user and the computer application of the preceding functions on the text. That is, all the preceding functions may produce interesting and pertinent results but may also be accessible in a format that is more or less cognitively transparent. Indeed, the various CARAT functions may produce strict and rigorous results but they are so large if not huge these results can not be cognitively process by a ordinary user. (Rosenblaum, 1992, Di Batista et all, 1999; Hearst 1994)

Configuration is the term we apply to a set of functions that translate the results obtained through configuration into a representation which is cognitively pertinent for the user-interpreter. Often, especially when the corpus is extremely large, configuration yields results which are too far-reaching and whose structure is inaccessible to the user. For example, a configuration could yield 75 pages of concordances for a request concerning a text. More often yet, using the voluminous information arising from the text, a user will wish to build a structured data base that will allow him to navigate through his or her results according to what interests him.

Here one needs functions that render the CARAT processing cognitively more helpful. We have called these function *configuration functions.* These functions will often come after a classification or a categorization. In one sense, they are also a categorization function but where another set of constraints are present: the cognitive understanding by the user. Here is a first list of such functions:

Listing: presenting the results in a vertical sequence.

Mapping: presenting result into a structured graph of some sort. (tree, maps. Etc) (Herman and all. 1999)

Visualisation: presentation of result re-injected into a type or other of iconic description. (Tile(tile, 3D space graphs, etc)

The four broad classes of operations - inscription, classification categorization and configuration- are important functions any computer-assisted text reading or analysis will have to carry out. It should be noted that these operations are functions which are defined here abstractly, independently of their computer instantiation.

The computer design: SATIM.

The preceding analytic and functional design allows the development of concrete computer architecture. It is here were the units of information, functions conceived in the preceding analytic and analysis chains are transformed into a computer application, Hence unifs become *data*, functions *become* modules, and analysis chains become complete applications, (sequences of modules). This is the base for the SATIM architecture for CARAT.

This architecture is a computer platform that implements the functional principles presented above. An anterior version of it was ALADIN (Seffah, Meunier, 1995) that offered an integrated and polyvalent platform for integrating the multiple operations requested by a user. But it still was a closed and proprietary; The SATIM approach goes farther in this vision. It offers a platform that allows to compose, test and experiment independent modules, have them communicate and construct well-formed sequences of modules.

More specifically, it offers three levels of construction specialized either for a conceiver, a researcher and a end user. We can conceptualize the SATIM Platform as presenting three levels of architecture that are called the *workshop*, the *laboratory* and the *application*.

1. The workshop.

This first level platform is an incubator and a repository for the various modules (functions) that are to be used in building a analysis chains. These modules are autonomous and independent. They even can have been built in various programming languages. This is a condition for maintenance and updating the workshop.

The SATIM workshop is in fact a special type of tools box. It is not so much a set of modules (functions) but a set of tools for working on the modules. These tools allow them to communicate their input and output in view of a specific research objective or processing.

Here SATIM relies on three computers programming approach: the object-oriented design, the multi agent and most of all the combinatorial functional design presented above. In its actual form, SATIM is in fact a huge relational data base managing a variety of existing modules (lemmatizer, matrix builders, classifiers, statistical packages etc) and apt at accepting future modules. This workshop is for the moment only accessible to expert programmers and its design is part of an ongoing software engineering project.

2. The laboratory.

The second level of the SATIM platform offers CARAT researcher tools allows him to explore, in a transparent manner various analysis chains, build out of the precedent tools box and repository. The user here needs not to be a computer programmer but he must be an expert in the field of a CARAT application. Through a ergonomic interface, this user chooses the modules he wishes to see functioning in one or other of the analysis chains. He is assured that these chains are syntactically well formed that is : they form a complex algorithm. Their semantic though (i.e., their meaning) depends specifically on the tasked aimed at. That is, even though the modules can be combined in a well form manner, they are not necessary semantically pertinent for one particular task.

In is actual form, two analysis chains have been built: Numexco and Grammexco. These two chains are classifications chains on texts. One works on words as basic units of information, the other on n-gramms. Other chains are being built as research goes on. Indexico for indexing, Ontologico for ontology maintenance. Thematico for thematic analysis. Categorico for automatic categorization of texts.

3. Applications

The third level of the SATIM architecture is for the end user. If after many tests and experiments a particular successful chain is finally accepted it can be wrapped up as an autonomous application, transparent to this end user and where only certain parameters are modifiable. (v. g. a particular natural language). It may have its own interface. And if a particular chain does not fit a specific goal objective and modules have to be changed, one must go back in the laboratory and experiments new chains .

With SATIM, some applications of specific chains have been applied to certain domains: Philosophy: (De Pasquale & Meunier 2002; Forest & Meunier 2000) , linguistic (Biskri & Meunier. 2002)

Conclusion.

Obviously, these three CARAT modelling are still very schematic and general. We nevertheless believe they give an insight into the process expert text readers and analysis applies on text, and help build in a systematic way a reliable and acceptable technology that clings to the concrete practices of the experts in the humanities and social sciences. This is aim of the SATIM approach. It is not mainly a technology but a complex modelling process for a software engineering of CARAT . We believe that these three levels of design help define better specifications for a modular architecture that can assist s the process in computer assisted reading and analysis of text.

References

Alexa, M. et Zuell, C. (1999). Commonalities, difference and limitations of text analysis software: the results of a review. ZUMA arbeitsbericht: Mannheim.

Barry, C. A. (1998.) « Choosing qualitative data analysis software: Atlas/ti and Nudist compared ». *Sociological Research Online*, vol. 3, no 3. www. socresonline. org. uk/socresonline/3/3/4. html.

Benveniste (Émile),(1966) *Problèmes de linguistique générale* (2 tomes), Gallimard, 1966-1974.

Biskri. I. Meunier, J. G, (2002) SATIM : *Système d'Analyse et de Traitement de l'Information Multidimensionnelle,* St Malo Les Journées internationales d'Analyse statistique des Données Textuelles.

Bradley John, Geoffrey Rockwell (1992) *Towards new Research Tools in Computer-Assisted Text Analysis* Presented at The Canadian Learned Societies Conference, June,

Brunet, E. (1986). *Méthodes quantitatives et informatiques dans l'étude des textes*. Paris : Champion.

Burton Orville Vernon (ed) (2002) Computing in the Social Sciences and Humanities Univ of Illinois Pr (Pro Ref); Book and CD-ROM edition (March),

Curry, B. H., Feys, R., 1958, *Combinatory logic*, Vol. I, North-Holland.

Church, A. (1941). *The Calculi of Lambda Conversion.* Princeton: Princeton University Press.

Ryan , Marie-Laure (Editor Cyberspace Textuality: Computer Technology and Literary Theory

Popping, R. Roel (2000) *Computer-assisted text analysis* London : Sage Publications, 2000

Damashek. M. (1989). "Gauging similarity with n-grams: language independent categorization of text. ". *Science* 267: 843-848

De Pasquale, Jean-Frédéric et Meunier, Jean-Guy, (2003) Categorisation techniques in computer assisted reading and analysis of text (CARAT) in the humanities », Proceeding of the ACH/ACLL Conference, Computer and the Humanities, Kluwer, Volume 37, No. 1, Février

Desclés, J. P. (1990), *Langages applicatifs, langues naturelles et cognition.* Paris: Hermès.

Biskri, I., Desclés, J. P., (1997), « Applicative and Combinatory Categorial Grammar (from syntax to functional semantics) », dans *Recent Advances in Natural Language Processing (selected Papers of RANLP 95)* Ed. Ruslan Mitkov & Nicolas Nicolov. John Benjamins Publishing Company, Numéro 136.

Fielding N. G. and R. M. Lee, (eds). (1991). *Using Computers in Qualitative Research*. Thousand Oaks, CA: Sage. ,

Forest, D. et Meunier, J. -G. (2000) "La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques" in Rahman, M. & Chappelier, J. -C. (ed.). *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles,* 9-11 mars 2000, EPFL, Lausanne, Suisse. Volume 1, pages 325 à 329.

Fortier, P. A. (2002). Prototype effect vs. rarity effect in literary style. In Louwerse, Max and Willie van Peer (eds.), *Thematics: Interdisciplinary Studies*, x, 448 pp. (pp. 397–405).

Glaser, B. G. et Strauss, A. L. (1967). *The discovery of grounded theory. Stategies for qualitative research.* Chicago: Adline.

Greenstein, Daniel I. A (2002) *Historian's Guide to Computing* (Oxford Guides to Computing for the Humanities.

Grosz Barbara J. and Candace L. Sidner. (1986) Attention, intentions, and the structure c discourse. *Computational Linguistics,* 12(3):175-204, July-September.

Hearst Marti A. . (1994a) *Context and Structure in Automated Full-Text Information Access***.** PhD thesis, University of California at Berkeley, 1994.

Hirschberg, Julia and Diane J. Litman. (1987) Now let's talk about *now:* Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL-87),* pages 163–171, Stanford, CA,

Hobbs Jerry R. (1990) *Literature and Cognition*. CSLI Lecture Notes Number 21, Stanford, CA: Cambridge University Press.

Hockey by Susan M. (1994) *Electronic Texts in the Humanities: Principles and Practice* Oxford University Press.

Ide, N. (1992). *Text Software Initiative*. ists. village. virginia. edu/lists_archive Humanist/v07/0031. html

Kastberg Sjöblom, M. et Brunet, E. (2000). « La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain ». In Rajman, M. et J. -C. Chappelier (dir. publ.). Actes des 5ièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT). Lausanne, 9-11 mars 2000, vol. 2, p. 457-466. Lausanne : EPFL.

Lévesque, G. (1998). *Analyse de système orientée-objet et génie logiciel, concepts, méthodes et application*. Montréal : Chenelière/McGraw-Hill.

Levesque, H. J. (1984). *Foundations of a Functional Approach to knowledge Representation*. 155-212.

Lancashire, Ian, John Bradley, Willard McCarty, Michael Stairs, and T. R. Wooldridge. *Using* TACT *with Electronic Texts: Text Analysis Computing Tools Version 2. 1*. New York: MLA, 1996. With CD-ROM.

Lusignan, S. (1985). *Quelques réflexions sur le statut épistémologique du texte électronique*. Computers and the humanities, vol. 19, p. 209-212.

Mann , William C. and Sandra A. Thompson.(1988) *Rhetorical structure theory: Toward a functional theory of text organization*. Text, 8(3):243—281,

Marcu, D. (2000) The theory and practive of discourse parsing and summarization. MIT.

Mayaffre, D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres.* Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939). Paris : Honoré Champion.

McKinnon, A. (1979). « Some conceptual ties in Descartes' "Meditations" ». Dialogue, vol. 18, p. 166-174.

McTait K. (1998) *A Survey of Corpus* Analysis Toolshttp: //www. limsi. fr/ Individu/mctait/Publications/Survey. pdf

Meunier J. G, Seffah, A., (1995) ALADIN, Un atelier orienté objet pour l'analyse et la lecture de Textes assistée par ordinaleur in S. Bolasco. L. Lebart, A. Salem Analisi Statistica dei Datin Testuali, vol II, CISU. Rome, 289-292, in vol 2 p 105. 112, vol 2, Meunier,J.G. F. Daoust, S. Rolland, Sato: A system for Automatic Content Analysis of Text, *Computer and the Humanities* 1976, vol XX.

Montague, R., (1972) Formal Philosophy/Selected Papers of Richard Montague, Thomason editor. Yale Univ.Press

Nielsen, J. (1986). "Online Documentation and Reaser Annotation*". International Conference on Work with Display Units*, Stockholm, 12-15 mai.

Pêcheux, M. (1972). *L'analyse autamatique du discours*. Paris : Maspéro.

Polanyi Livia. *A formal model of the structure of discourse*. Journal of Pragmatics, 12:601—638, 1988.

Popping, R. (2000) *Computer-assisted text analysis* Roel Popping. London : Sage Publications, New technologies for social research

Rada, R. (1991). *From Text to Expert text*. Mc Graw Hill.

Ram, A. Moorman, K (eds). (1999) *Understanding Language Understanding,, Computational Models of Reading*, MIT Press

Rastier, F. (2001). *Arts et sciences du texte*. Paris : Presses Universitaires de France.

Richards, T. J. and L. Richards. (1994) Using computers in qualitative analysis. In Denzin, N. and Lincoln, Y. (eds), *Handbook of Qualitative Research*, pp. 445-462, Berkeley, CA: Sage.

Ryan M. L. (1999). "Introduction" and "Cyberspace, Virtuality and the Text. " Both in *Cyberspace Textuality: Computer Technology and Literary Theory*. Ed. Marie-Laure Ryan. Indiana University Press. 1-29 and 78-107.

Salton Gerard, Amit Singhal, Chris Buckley, and Mandar Mitra. (1995) *Automatic text decomposition using text segments and text themes.* Technical Report TR-95-1555, Department of Computer Science, Cornell University.

Salton Gerard and James Allan. (1995) *Selective text utilization and text traversal. International Journal of Human-Computer* Studies, 43:483-497,

Sinclair, S. (2002 « Humanities Computing Resources: A Unified Gateway and Platform ». COCH/COSH 2002. Toronto: University of Toronto. 26-28 mai 2002.

Sloman Aaron (1978) *The computer revolution in Philosophy* : Philosophy, science and models of mind. Harvester Press and Humanities Press

Sowa, J. F. (1991). *Principles of Semantic Networks*. San Mateo: Morgan Kaufman. UML by Example

Unsworth, J. (2000). « Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? » part of a *symposium on Humanities Computing: formal methods, experimental practice.* London : King's College. May 13, 2000.

Shaumyan, S. (1987*). A Semiotic Theory of Language.* Bloomington Indianapolis: Indiana University Press.

Schönfinkel, M.(1924). `Uber die Bausteine der mathematischen'. *Mathematische Annalen*, 92, 305-316.

Computer-assisted phonetic analysis of English poetry

Marc Plamondon, University of Toronto

I will discuss and demonstrate three ways computers can aid the interpretive study of literature. The first is well known: computers can do for us some things we can already do for ourselves, but in a fraction of the time. The second receives less attention: computers can aid us in learning or refining particular interpretive skills. The third is more controversial: computers can aid us to quantify our analysis of a text. The quantification of beauty is the underlying concern in my paper. While the process of assigning a numerical value to an aesthetic object is not necessary to developing an appreciation of that object or even making a critical statement about that object, it may prove helpful to explaining our appreciation and thereby helpful to making critical statements. I shall discuss the use of quantifying one aspect of poetry — its sound — as one step towards the quantification of poetic beauty.

The phonetic analysis of poetry is not easily accessible to most students — at any level – of poetry. Once one has learned the terminology, the symbols, and the linguistic rules, one is left facing a large amount of raw data contained in a poem. A four-line poem is a collection of about one hundred phonemes. Browning's "My Last Duchess," a moderate-length poem of 53 lines, has 1551 phonemes (if one includes optional sounds). What is the weary student to do when faced with a poem of 1000 lines or more?

My computer application will take a poem, convert it into a series of phonemes, and display it as such for the user. The user can then click on any phoneme to have all of its occurrences in the poem highlighted. Various combinations can be selected with user-defined colour schemes. In doing so, the face of the poem changes significantly: by allowing a user to see a poem with some sounds highlighted in one colour and some other sounds highlighted in another colour, the application allows the user to see the sounds of the poem. A visual representation of poetic sound makes (I believe) the sound of the poem more accessible to students of poetry: they can see representations of auditory patterns more easily than they can hear those patterns. But once they can see the patterns, they can then listen for them. The computer application should help students move beyond merely hearing rhyme and assonance (when they are heard at all).

The application can assemble statistics about the phonetic makeup of a poem. The number of times a certain phoneme occurs in every line of the poem can be calculated, including as a percentage of the sum of all phonemes in the line or of some other sum. These numbers can then be represented visually as shades of a certain colour: the more frequently a particular sound occurs in a line, the brighter the colour of the line. This becomes more interesting when groups of phonemes are represented by the same colour:

the relative number of plosive consonants in a line can be represented visually by shades of the colour red. Even more fascinating is when two or three collections of phonemes are represented by their own colour and the combination is represented visually. For example, the relative occurrence of plosives, fricatives, and approximants in a line of poetry can be represented by the colours red, green, and blue. The result is an RGB value: a colour that represents the relative occurrence of the combination of the three groups of phonemes. A user can then easily see whether a line of poetry has a greater proportion of one group of consonant sounds or another and make interpretive statements based on the results. Aside from demonstrating this feature, I shall show some of the more interesting results I have found using it: including how semantically different but phonetically equivalent lines echo each other.

A similar RGB number can be used to represent a poem as a whole in order to compare one poem against another. This can be used to compare a group of poems by the same author: indeed, not only can one assign a colour to a line or a poem, one can assign a colour to a poet as the average of that poet's poem colours. These numbers can then be used to compare poets and poetic movements: did the Romantics or the Victorians use a greater proportion of fricative consonants? They can be used to help answer the century-and-a-half-long debate: which is the more musical poet, Tennyson or Browning? I shall discuss the value and limitations of this technique of using computer-generated statistics to answer questions of poetic aesthetics.

The application at this stage is primarily concerned with the visual representation of a poem: colouring its sounds. A further development of the application is phonetic pattern recognition. The identification of phonetic patterns in poetry is, up to a certain point, similar to the identification of metrical patterns. I shall discuss the extent to which a successful algorithm for metrical analysis can be applied to phonetic analysis, and the use and limitations of that application.

In Praise of Pattern

Stephen Ramsay, University of Georgia

A few years ago, Hugo Craig expressed quite plainly a sentiment that I suspect haunts many of those who work with computer-assisted text analysis:

> The leap from frequencies to meanings must always be a risky one. The interpreter who is tempted to speculate about the world-view or psychology of a writer, based on quantitative findings, presents an easy target for dismissive critique (Fish, 1973). On the other hand, staying within the safe confines of the statistical results themselves means that one is soon left with only banalities or tautologies. Lower-level features are easy to count but impossible to interpret in terms of style; counting images or other high-level structures brings problems of excessive intervention at the categorization stage, and thus unreliability and circularity (Van Peer, 1989). (Craig 1999)

Not all practitioners are haunted by this sense of peril and self doubt. Computational linguists, who are naturally loath to extrapolate world-views from columns of morphological data, are for the most part unimperiled by banalities and tautologies. They use machines adept at isolating low-level structures to study low-level phenomenon, and the activity has been yielding useful research for decades. Digital librarians are likewise untroubled by the possibility of excessive intervention from their search engines and rendering tools. They endeavor to preserve and provide access to digital texts, and their efforts at doing so have been successful by any measure. Even designers of traditional text analysis tools like collators, concordancers, and word-frequency generators can, for the most part, work without fear of unreliability and circularity. Interpreting those results is somebody else's peril.

But those of us who use computers in the context of what I would like to call the "hard humanities"—philosophy, literary criticism, exegetical historiography, aesthetic evaluation—are extremely troubled by these matters, and should be. The hard humanities are not, as Gadamer noted, "concerned primarily with amassing verified knowledge, such as would satisfy the methodological ideal of science" (Gadamer xxi), but with truth and knowledge that emerges amidst the tectonic forces of textuality, language, and connotation. We are not "tempted to speculate," but required to do so.

Susan Hockey was right when she wrote that computers are "best at finding features or patterns within a literary work and counting occurrences of those features," and the observation is generally true across most texts of interest to humanists literary or otherwise (Hockey 66). Finding and counting might indeed be the two most fundamental activities of computing as such. Those of us who propose to use computers in the primarily interpretive discourses of the humanities, though, need to discover how the low-level process of finding and counting relates to the high-level practice of interpretation. If it is matter of finding words and counting them, then Craig is surely right. However useful this might be for certain areas of humanistic inquiry (like corpus linguistics), it seldom helps us to get at the sort of problems with which historians, philosophers, and literary exegetes grapple. Part of the problem, however, has been our insistence that text analysis prove or verify the existence of low-level patterns, instead of using low-level patterns to help us locate the broader patterns upon which interpretive work depends.

Computers cannot (at least at the moment) enter humanistic discourse as fellow exegetes, but they can do a lot more to facilitate the interpretive process. Data mining offers us a way to bridge the gap between the quotidian search for low-level features and the higher-level patterns that serve to prompt interpretive insight. The essential move of data mining is the search for patterns among broad, at times apparently unrelated low-level features. As such, it resembles more closely the act of critical reading, which is at heart concerned with the less-than-obvious patterns that tend to escape more passive readings.

This paper presents some preliminary results from the use of data mining techniques to facilitate the study of dramatic structure in Shakespeare. begin by demonstrating StageGraph—a program that can read in a simple XML representation of the character-scene matrix in a Shakespeare play and generate directed graphs of the relationships between characters and scenes. The resulting visualizations are not without interest; they show us broad differences among the size and structure of Shakespeare's plays that would be difficult to see in absence of such an aid. However, StageGraph can also generate tables of graph properties that correspond roughly to the standard mathematical features used to describe graphs. This latter function gives us a set of extremely low-level data points: number of nodes, number of edges, average degree, chromatic numbers, cycles, diameters, radii, and so forth. None of them are of themselves particularly enlightening: *Antony and Cleopatra* has more scenes that *A Midsummer Night's Dream*; *As You Like It* has the most cycles of any of Shakespeare's plays, etc. However, when we use data mining techniques to find patterns among these low-level features, certain very suggestive patterns begin to emerge.

Decision-tree generation and naive Bayesian clustering of the data points demonstrate that the strongest distinguishing feature among the graph structures of the plays seems to be the incidence of single scenes (scenes that are used only once). The comedies, by and large, have few such singletons, while the histories use them frequently. Attempts to plot all of the plays using similarity measures, moreover, reveal several anomalies—the most interesting being, perhaps, that Othello and Romeo and Juliet appear to be tragedies with the structure of comedy.

The computer does not say why. Nor does it assure us that the data points we have chosen are the only ones likely to yield interesting results. Instead, it mines through a set of seemingly innocuous data values and prompts us to look where we weren't looking before. I argue that this change of vantage point is more valuable and more useful for interpretive inquiry than any results the procedures might yield—in part because this change of vision lies at the heart of critical reading itself.

Works Cited

Craig, Hugh. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?." *Literary and Linguistic Computing* 14 (1999): 103–13.

Gadamer, Hans-Georg. *Truth and Method*. Trans. Joel Weinsheimer and Donald G. Marshall. New York: Continuum, 1996.

Hockey, Susan. *Electronic Texts in the Humanities*. Oxford: Oxford UP, 2000.

Lightweight Morphology: A Methodology for Improving Text Search

Mikaël Roussillon & Bradford G. Nickerson, Faculty of Computer Science, University of New Brunswick
Stephen Green & William A. Woods, Sun Microsystems Laboratories

Lightweight Morphology is a new approach to morphological analysis, creating morphological variants from sets of rules. The rules are intuitive to define and at the same time, offering expressiveness and control. We defined a grammar for Lightweight Morphology. We defined how to generate English and French morphological variants using the grammar. French Language specification required 526 rules, 41 rule sets and 16,842 exception table words, while the English language specification took 123 rules, 17 rule sets and 2,589 exception table words. English and French Lightweight Morphologies were compared with two other techniques extending queries on a collection of 533 documents from the aligned Hansard of the 36th parliament (1997) of Canada. A differential recall comparison among the techniques showed that Lightweight Morphology has more queries (average of 3.9 times more) retrieving fewer irrelevant document for both English and French. The French Lightweight Morphology has more queries (average of 2.5 times more) retrieving more relevant documents.

1 Introduction

Interpretation of a text can be assisted by knowing the location and content of ‘like’ phrases and words using morphological variants. The challenge of using morphological variants is how to generate them. One can (a) have a complete lexicon for each language that includes all morphological variants, or (b) encode morphotactics as rules, so as to generate the variants of recognized morphological patterns [5, 8]. For example, in French, the ending ’erait’ as in ‘aimerait’ identifies the conditional 3rd person singular of a first group verb, and we can produce the morphological variants corresponding to the first group verb paradigm, such as ‘aimer’, ‘aimes’, ‘aimions’ and many more.

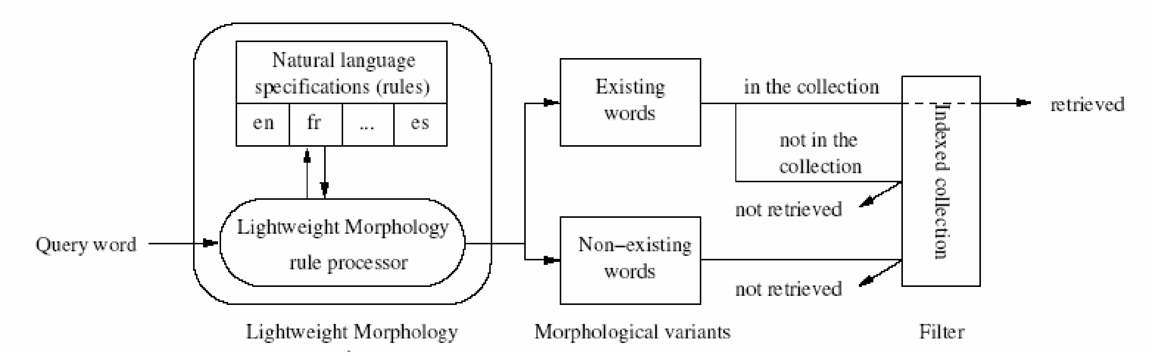


Figure 1: Lightweight Morphology variants production with rule specifications and filtering against the collection (en stands for English, fr for French, es for Spanish).

2 Lightweight Morphology

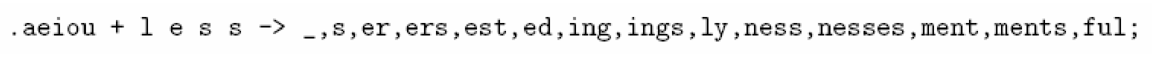
Lightweight Morphology is a methodology that follows the second approach. This approach is space efficient and avoids the complexity of establishing a complete lexicon of all morphological variants. As a preprocessing step, Lightweight Morphology tries to expand the word into multiple morphological variants, therefore adding information to the word (see Figure 1). Lightweight Morphology consists of three components which are (1) a set of pattern-matching rules enclosed in rule sets that will produce morphological variants, (2) an escape mechanism to write rules in Java™ programming language and (3) an exception table to handle exceptions that exist in a language. Lightweight Morphology has a modular approach that enables one to define morphologies for different languages or to introduce ew approaches on the way the morphological variants are to be produced.

2.1 Pattern-matching rules and rule sets

A pattern-matching rule is made of the following elements:

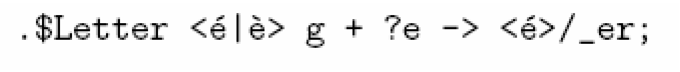
1. On the left hand side of the rule, a pattern the word has to match in order to apply the rule. The pattern is defined with regular expressions, and some mechanism is provided to interact with the right hand side.
2. On the right hand side of the rules, a list of morphological variations that are to be applied in order to obtain the variants.
3. The left hand side is separated from the right hand side with the '->' production symbol.

The operators of the pattern-matching rules allow to handle different kinds of affixes, like circumfixes and infixes, either by removal from the input word or by addition to the produced variants .Diphthongs and spelling modifications can also be handled. Figure 2 presents a simple example of a pattern-matching rule for English processing words ending with a vowel followed by les. The word is stripped from *less* and the list of morphological variation from the right hand side of the rule are appended to form the morphological variants ('_ ' stands for the empty morphological variation). Figure 3 shows a more complex example to create the infinitive forms for French verbs like *assiéger*. It is supposed that the input to this rule is a verb stripped from its ending.



```
.aeiou + l e s s -> _,s,er,ers,est,ed,ing,ings,ly,ness,nesses,ment,ments,ful;
```

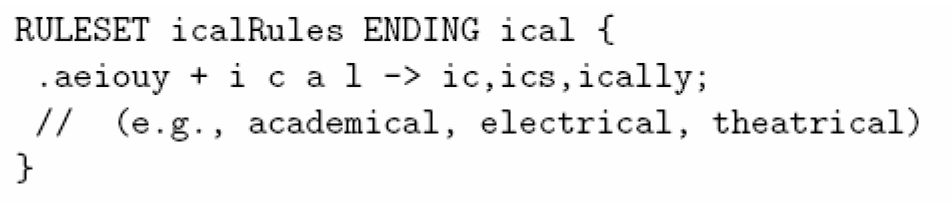
Figure 2: Pattern-matching rule example for English words ending with less.



```
.$Letter <é|è> g + ?e -> <é>/_er;
```

Figure 3: Pattern-matching rule example for French infinitive form creation.

The pattern-matching rules are themselves ordered within a rule set. The first rule that will be successful in the rule set (i.e. the pattern matches the word and the right hand side produces variants) will return the morphological variants obtained and no more rules will be tried. Different kind of rule sets can be used. The first one is the default rule set, the one that will be first tried when performing Lightweight Morphology. The ending rule set offers a different behaviour. It encloses rules that will apply on words having the same ending. Therefore, if an input word has an ending defined by an ending rule set, this specific rule set will be tried instead of the default rule set. Figure 4 presents an ending rule set containing the only rule processing English words ending with ical. The third kind of rule sets are Java™ programming language rule sets and will be explained in the next section. The fourth kind is a rule set whose sole purpose is to group a set of pattern-matching rules.



```
RULESET icalRules ENDING ical {
 .aeiouy + i c a l -> ic,ics,ically;
 //  (e.g., academical, electrical, theatrical)
}
```

Figure 4: Example of ending rule set enclosing a rule for words ending in ical.

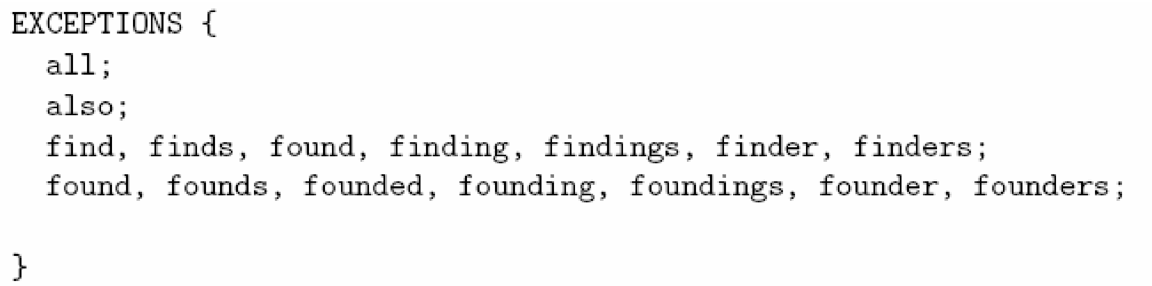
Each rule set can be called from a pattern-matching rule, the word matched and modified by the pattern-matching rule being the input word to the rule set. The usefulness of rule sets is therefore to group similar processing, either because of their ending which represents a certain morphological feature (e.g. the *-s* ending for plural), or because of the same problem they solve (e.g. creating from a specific kind of verb the present forms).

2.2 User-defined rules using Java™ programming language

Sometimes, pattern-matching is not enough to describe morphotactics and create morphological variants. For those cases, Lightweight Morphology offers the possibility to write rules directly in Java™ programming language to create user-defined rules. These rules behave much like rule sets since pattern-matching rules can call user-defined rules and user-defined rules can call rule sets. Different user-defined rules can coexist by creating different Java™ programming language rule sets.

2.3 Exception table

Every language has exceptions or irregularities, and some are so exceptional that they cannot be encoded as rules. Most of the time, those exceptions are intensively used words that earned a very irregular pattern over the centuries. Classical examples that apply to many languages are 'to be' and 'to go'. Lightweight Morphology provides a mechanism to write this kind of exception. The exception table can be seen as a kind of lexicon where morphological inflections and derivations of a word a grouped together as an entry. Figure 5 presents a small portion of the exception table for English words.



```
EXCEPTIONS {
  all;
  also;
  find, finds, found, finding, findings, finder, finders;
  found, founds, founded, founding, foundings, founder, founders;

}
```

Figure 5: Exception table for some English words.

Before processing an input word with rules, Lightweight Morphology first checks if the word appears in the exception table. If so, all entries containing the word are returned and no rules are tried.

The exception table can also be used for words where rules produce incorrect morphological variations (morphological variations that are real words but are not related to the input word), or for words that do not require processing by rules, like adverbs.

3 Lightweight Morphology specification

The essential work of Lightweight Morphology is to define pattern-matching rules for natural languages. Depending on the language and the person creating the specification, various approaches can be taken to construct a Lightweight Morphology specification a natural language. We present two approaches, one for English and one for French. Use of grammatical resources, like *Bescherelle* [1], *Bled* [2] or *Grevisse* [4] for French, is strongly suggested.

3.1 English

The English Lightweight Morphology has 123 rules, 17 rule sets and 2,589 exception table words. The rule sets are ending rule sets, except for the default rule set. The processing is therefore performed depending on the ending of the input word. Example of endings processed are *-s*, *-less* or *-ing*, handling noun and verbs inflection and some derivations. The default rule set only processes endings that are not defined as an ending rule set. Heavy use of rule set calls is made, for example in some pattern-matching rules, words stripped from *-s* will go again through the rules as a new input word and create new morphological variants.

The exception table is an important part of the English Lightweight Morphology. It encodes irregular verbs (e.g. *find*) or words that fail to be processed by the rules (e.g. *firmament*).

On the set of 200 English words processed by the English Lightweight Morphology, 4% produced irrelevant variants (e.g. *billy* from *bill*) and 16% are did not produce one or more important variant (e.g. *legislate* is not produced from *legislation* because of an unhandled derivation).

3.2 French

The French Lightweight Morphology has 526 rules, 41 rule sets and 16,842 exception table words. All rule sets are normal rule sets, except for the default rule set. Each normal rule set will perform a specific grammatical feature. For example, a rule set is used to process masculine adjectives and nouns, while another rule set is used to create the infinitive form for 1st group verbs. The default rule set dispatches the words to the appropriate rule set according to a pattern.

Support for derivation for the suffixes *-age, -ance, -ment, -eur, -ion,* and *-able* is provided creating morphological variants from and to these suffixes.

The exception table is extensively used to encode irregular verbs(e.g. *aller*, to go) and irregular form of adjectives (e.g. *beau*, beautiful) and nouns (e.g. *carnaval*, carnaval).

The pattern-matching rules, along with the exception table, are supposed to handle inflections for verbs, nouns and adjectives. On a set of 200 French words, 5% produced incorrect variants (e.g. *paye* from *pays*), and 7,5% did not produce one or more important morphological variant (e.g.*rapporter* is not produced from *rapport*).

4 Testing

We tested the Lightweight Morphology approach with 200 English query words and 200 French query words in the aligned Hansard of the 36th parliament of Canada [3] which consists of 533 documents (19,999,604 tokens for the English version, 22,801,063 tokens for the French version). The English version contains 55,323 terms and the French one 76,031 terms. By term we mean a token that does not contain a cipher.

We compared Lightweight Morphology with two other search approaches: (a) stemming and (b) wildcards. For example the word 'academic' produces 'academics', 'academical', 'academically' for Lightweight Morphology and 'academ' for the Porter stemmer. A good wildcard query would be 'academ*' or 'academic*'. We included the number of documents retrieved with exact query (no preprocessing) as a reference measure, because it is common among search engines.

The Porter stemmer [6] algorithm is used for English, a stemmer created by Martin Porter [7] is used for French. For the French Lightweight Morphology, we produced two sets of rules: (1) one that performs inflections and derivations and (2) one that only performs inflections.

For each language, 100 query words are randomly selected from an Ispell dictionary. If a word does not have a hit in the collection with one of the approaches, it is discarded. The remaining 100 query words are frequency selected from the Hansard. We sorted the words of the collection by frequency and selected the first 100 meaningful words. We consider that a meaningful word is a word that is an uncommon word. Common words can be an article (e.g. *the, a*), a common verb (*be, go*) or any word that would not make much sense in a query as a single word (e.g. *however, first*).

To evaluate Lightweight Morphology against the other approaches, we used a qualitative measure known as differential recall. In differential recall, we compare two methods A and B and calculate:

- $A \cap B$ — The number of relevant documents found with variants from both A and B;
- $A - B$ (resp. B− A) — The number of relevant documents found by variants from A(resp. B) but not from B(resp. A).

We decided that our relevance criterion would be the correctness of the variants found (e.g. for the query 'tear', we consider 'torn' to be relevant but not 'tearmann'), without consideration for the context (e.g. we consider 'tear' to be relevant both in the context of a teardrop and of something that is torn).

For the 200 words of each language, we manually classified the variants produced by the three approaches as either relevant or irrelevant according to our relevance criterion.

5 Results

The following abbreviations will be used: **LM** (Lightweight Morphology), **LM'** (Lightweight Morphology without derivation processing), **S** (Stemmer), **W** (Wildcard) and **EQ** (Exact Query).

5.1 Relevance classification and differential recall

Some results for three selected English query words are presented in Tables 1, 2 and 3 .

The results present a good sample of words retrieved by the approaches. The relevance/irrelevance classification is not perfect in part because of the collection that contains typos, spelling errors, neologisms and foreign words (many French words in the English Hansard and vice-versa). Judging a word to be relevant is often subjective, and our classification can be contested for some queries. For those few contestable classifications, we believe the impact on the differential recall is minor.

5.2 Differential recall win-lose score

In order to compare differential recall among all the words and all the approaches, we introduced a win-lose score, based on the 'difference' score ($A- B$ and $B- A$) from differential recall measures. For the relevance criterion, if $A- B > B- A$ for a query, A is awarded 1 point (A is retrieving more relevant document than B on the query). For the irrelevance criterion, if $A- B > B- A$ for a query, B is awarded 1 point (B is retrieving less irrelevant documents than A). The results of the win-lose scoring is given in Table 4 for the randomly selected words and in Table 5 for the frequency selected words.

Table 1: Variants found with Lightweight Morphology, Stemming and Wildcard.

Query	Approach	Example of English variants found in Hansard (number of documents found with this variant, relevance)
artist(85) (artist*)	LM	artist(85, rel), artists(136, rel)
	S	artist(85, rel), artists(136, rel), artistes(2, rel), artistic(46, rel)
	W	artist(85, rel), artists(136, rel), artistes(2, rel), artistic(46, rel), artistically(2, rel), artistique(1, rel), artistry(4, rel)
leaf(99) (leaf*)	LM	leaf(99, rel), leafs(21, rel), leafing(2, rel), leaves(299, rel)
	S	leaf(99, rel), leafs(21, rel), leafing(2, rel)
	W	leaf(99, rel), leafs(21, rel), leafing(2, rel), leaflet(3, irrel), leaflets(5, irrel)
tear(85) (tear*)	LM	tear(85, rel), tears(100, rel), tore(17, rel), torn(97, rel), tearing(40, rel)
	S	tear(85, rel), tears(100, rel), teared(1, rel), tearful(2, rel), tearing(40, rel)
	W	tear(85, rel), tears(100, rel), teared(1, rel), tearful(2, rel), tearfully(1, rel), tearing(40, rel), tearmann(1, irrel)

Table 2: Differential recall of Lightweight Morphology vs. Stemming for English.

Query Word	LM – S	LM ∩ S	S – LM
artist	0	153	16
leaf	223	110	0
tear	55	182	1

Table 3: Differential recall of Lightweight Morphology vs. Wildcard for English.

Query Word	LM – W	LM ∩ W	W – LM
artist (artist*)	0	153	19
leaf (leaf*)	223	110	0
tear (tear*)	55	182	1

As documents can be retrieved by more than one variant, results from Table 1 do not necessarily translate to the numbers in Tables 2 and 3.

Table 4: Differential recall WinLose scores for 100 randomly selected words.

Language	Criterion	LM S	LM' S	LM W	LM EQ	S EQ
English	Relevant	24 - 24	N - A	8 - 44	82 - 0	81 - 0
English	Irrelevant	4 - 1	N - A	39 - 0	0 - 1	0 - 4
French	Relevant	35 - 13	25 - 36	11 - 33	90 - 0	93 - 0
French	Irrelevant	5 - 2	5 - 1	26 - 1	0 - 2	0 - 5

Table 5: Differential recall WinLose scores for 100 frequency selected words.

Language	Criterion	LM S	LM' S	LM W	LM EQ	S EQ
English	Relevant	11 - 17	N - A	8 - 33	71 - 0	73 - 0
English	Irrelevant	21 - 4	N - A	63 - 2	0 - 9	0 - 25
French	Relevant	17 - 7	12 - 13	23 - 14	73 - 0	70 - 0
French	Irrelevant	20 - 5	19 - 3	50 - 7	0 - 11	0 - 23

First, we can see the usefulness of morphological analysis by comparing LM to EQ and to W queries. Compared to EQ, LM provides more relevant documents on70 to 90 queries but introduces irrelevant documents on 1 to 11 queries. Most of the time, performing morphological analysis on terms will introduce relevant documents, and in a few cases, irrelevant documents.

When comparing to W, we can see the advantage of performing morphological analysis: the user does not need to think about the query. Wildcard queries generate comparatively more irrelevant words. For example, *pity* creates the query *pit** so it will retrieve unrelated words like *pitbull*. It is easy to often retrieve more relevant documents with a straightforward wildcard query; it is also easy to retrieve more irrelevant documents. Morphological analysis has a strong advantage since the user does not need to think about putting wildcard operators at the right place to ensure retrieval of many related and few unrelated documents.

Compared to S, English LM is better, not always by retrieving more relevant documents but by always retrieving fewer irrelevant documents. We can see that for English, LM has a small advantage over S on the first query set by retrieving fewer irrelevant documents in 4 queries and more in only . On the second query set, S retrieves more relevant documents on 17 queries (11 for LM). LM retrieves fewer irrelevant documents: in 21 queries, S retrieves more irrelevant documents and retrieves less irrelevant documents in only 4 queries. The Porter Stemmer for example, found the stem *intern* from *international* allowing variants such as *internalize* or *interned*. This illustrates that English LM can improve information retrieval systems by decreasing the number of irrelevant documents retrieved on a per query basis.

French LM performs better than S in every case, getting more queries retrieving more relevant documents and more queries retrieving fewer irrelevant documents. French Lightweight Morphology does not find verbs from nouns ending with a consonant (e.g. *accorder* from *accord* or *travailler* from *travail*) reducing its relevant retrieval efficiency.

Processing derivations in French is important. Ignoring the suffix derivations decreases the relevant documents retrieved, making LM' less efficient than S for returning relevant documents. The number of queries retrieving fewer irrelevant documents is almost unchanged, illustrating that processing derivation does not appear to include spurious variants.

6 Conclusion

Lightweight Morphology is a new approach on morphological analysis offering more control than stemming over the variants retrieved. The way Lightweight Morphology is defined is intuitive compared to the way stemming algorithms are defined. The consequence is that English Lightweight Morphology will outperform stemming by retrieving less irrelevant documents on a pr query basis. Stemming has a small advantage at retrieving more relevant documents on a per query basis, essentially because the English Lightweight Morphology does not process some derivations. The French Lightweight Morphology was superior in both having more queries retrieving more relevant documents and fewer irrelevant documents.

Future research for Lightweight Morphology involves creating specifications for other natural languages and seeing where to improve the limits of pattern-matching rules. Java™ programming language rules should be reworked and replaced by a specific language performing advanced string manipulation. Finally, other tests should be performed on other collections and with other query words.

7 Acknowledgements

Sun Microsystems Inc. is gratefully acknowledged for providing funding for this research. Thanks are also due to the University of New Brunswick Faculty of Computer Science for their support. Experiments for this research were carried out on equipment established with funding from the Canada Foundation for Innovation (TAPoR project).

References

[1] *Bescherelle : La Conjugaison pour tous*. Hatier, Paris, 1997.

[2] E. Bled and O. Bled. *Bled* : *orthographe-grammaire.* Hachette, Paris, 2003.

[3] Ulrich Germann. Aligned Hansards of the 36th parliament of Canada release 2001 - 1a (proceedings from September 25, 1997). Resources available at http://www.isi.edu/ naturallanguage/download/hansard/.

[4] M. Grevisse. *Le Bon Usage: Grammaire française avec des remarques sur la langue française d'aujourd'hui.* Duculot, Paris, 11ème edition, 1980.

[5] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Linguistic Analysis, pages 191–202, 1993.

[6] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[7] M.F. Porter. French stemmer. http://snowball.tartarus.org/french/stemmer.html.

[8] W. A. Woods. Aggressive morphology for robust lexical coverage. Technical Report TR9982, Sun Microsystems, 1999.

The Markup Analysis Engine: A new role for content analysis in digital content development

Paul Scifleet & Concepcion S. Wilson, School of Information Systems, Technology & Management, The University of New South Wales

Although markup languages are now widely deployed to identify and define the structural and contextual characteristics of electronic documents, to date very little work as been undertaken to evaluate how the encoded elements of markup language vocabularies are being used by human authors and systems developers in practice.

While the benefits of representing complex document structures through descriptive encoding are widely acknowledged, it is increasingly apparent that adopting encoding schemas presents organisations with documentary forms that are often difficult to reconcile with existing information management processes. The rapidly changing information environment in which projects have occurred has often placed the experience gained from project to project well ahead of the development of shared methodologies for understanding, adopting and implementing markup language vocabularies.

Although encoding guidelines exist for most metadata schemas, it is increasingly apparent that providing organizations with the apparatus to construct digital objects does not resolve issues of assigning 'meaning' to individual objects.

In the maelstrom of innovations that is resulting from SGML and XML technologies, this question of 'how are markup languages being used?' is least clearly understood. As a result any overview of practice over the last decade is hard to provide and largely under theorised.

Our study addresses this problem by investigating and contributing to the development of methods for interpreting and understanding the markup that is applied to digital objects.

The analysis of encoded documents is being supported by the development of analytical software that reports on markup usage (i.e. the occurrence of tags used) in batches of encoded documents. By interpreting and reporting on the markup applied within documents it should be possible to comment on the implementation of markup languages and metadata schemas.

While the availability of software capable of parsing and interpreting encoded documents provides the opportunity for a large amount of quantitative data analysis, the challenge for markup analysis is to move beyond 'counting tags'. Understanding the phenomena of changing human communication that is presented in the widespread categorisation of content now underway across many different types of organisations requires an evaluation of how systems developers and the (human) encoders of digital documents are applying markup in practice. Our study addresses this problem by investigating patterns of markup usage in organisational context. We propose that making sense of the digital content object is reliant upon understanding of the object and its properties as they eventuate through commonly identifiable conceptual states in the development and implementation of markup languages. The paper presents a model for the analysis of content through a sequence of transitory states that is based on IFLA's Functional Requirements for Bibliographic Records.

This research takes its first steps towards the development of new applications for content analysis in digital content development. It is envisioned that over time, the development of analytical methods and quantitative procedures for markup analysis will contribute to:

- managing changing standards (e.g. identifying redundant and changing elements within documents);
- comparing, synchronising and merging different information sets;
- educating and training users involved in the design and development of markup based systems; and
- supporting research activities specific to organizational content or academic enquiry.

The research paper will focus on the establishment of requirements for content analysis in the context of the wider project and demonstrate the first version of markup analysis software and markup usage reports from the project. The demonstration will include documents encoded using the Text Encoding Initiative (TEI).

The Face of Meaning: A Short Presentation

Gary W. Shawver & Oliver Kennedy, New York University

Introduction

This short presentation is actually two shorter presentations. In the first part I will be talking about some of the theoretical assumption underlying the design of software that will be discussed and demonstrated in the second part by my colleague, Oliver Kennedy, one of our group's programmers. I first saw the desirability of such software while doing my doctoral dissertation, a computer-aided text analysis of the semantics of “story” and “tale" in Chaucer. It was clear that the lists generated by traditional texts analysis tools like TACT were both tedious to examine and presented information in a serial fashion not suitable to data that were in some sense relational. It was also clear that attempts to reformat the output of such “listware” into forms that more truly represented the nature of the data is laborious. Of course, a certain theoretical understanding of meaning, which I will briefly outline below, brought about this practical need.

Theory

> But if both meanings, or all of them. . . remain ambiguous after the faith has been consulted, then it is necessary to examine the context of the preceding and following parts of the ambiguous place, so that we may determine which of the meanings among those which suggest themselves it would allow to be consistent (Augustine, 2.2).
>
> The value of the chess pieces depends upon their position upon the chess board, just as in the language each term has its value through its contrasts with all the other terms (Saussure 88).
>
> For a large class of cases—though not for all— in which we employ the word ‘meaning’ it can be defined thus: the meaning of a word is its use in the language (Wittgenstein 20e).
>
> You shall know a word by the company it keeps! (Firth, “A Synopsis of Linguistic Theory, 1930-55” 179).

We assume in the design of this software a view of language that sees the meaning of a word as inextricably linked to its contexts, a view that is in some ways foreshadowed in Augustine’s advice to readers of Scripture to resolve semantic ambiguities by examining the context of a passage (3.2.2). To some extent, our software’s function and design is also informed by a Saussurean structuralist semantics that holds that a word’s meaning is “determined by the [horizontal] paradigmatic and [vertical] syntagmatic relations” between that word and others in a system of language (Lyons, *Semantics* I 268).

This is illustrated in Saussure's example of the game of chess in which the value of individual pieces is not the result of any intrinsic qualities but rather of their relation to other pieces (Saussure 108, 87 ff.). Of course, the method of deducing what a word named for Augustine becomes a mode of meaning for Wittgenstein, and while Saussure's focus is upon a word's place within a language system, that of Wittgenstein is upon a word's place within instances of discourse.

J. R. Firth was perhaps the first linguist to give Wittgenstein's statement serious consideration when he proposed that one could define a word by looking at its linguistic context, or "the mere word accompaniment, the other word–material in which [a word is] most commonly or most characteristically embedded" (Firth, "A Synopsis" 180). Seeking to differentiate this from "context of situation," Firth called it collocation ("Modes of Meaning" 195). In a sense, this study of collocation, the linguistic material or "linguistic context" within which a word commonly occurs, involves the mapping of pathways leading to and from each word (Palmer 76). Firth's idea of collocation, first presented in 1951, introduced a new type of word meaning, but would have to await the introduction of computers and electronic texts before it could be used to study texts of any significant size (Berry-Rogghe 103). The collection and classification of collocates require the computer's unique ability to perform repetitive tasks quickly and efficiently. Software like *TACT* brought this ability to desktop PC users and introduced a new set of challenges, one of which was how to represent the data gathered by such tools.

> A multitude of familiar paths lead off from these words in every direction. (Wittgenstein 143e, 144e)
>
> During the initial phases of memory encoding, it is as if one is preserving the shape of a letter by walking a particular route on a lawn. The pattern is dynamic, and is only evident in the way one moves. After a period of time, however, the grass will wear through, creating a dirt pathway. At that point, one can stop walking; the information is preserved structurally (Kosslyn & Koenig, 351)

The significance of a collocate lies in two things: repetition and syntagmatic relation. This brings us to the focus of our software, phrasal repetends, a subclass of collocation. Phrasal repetends fall into two categories: "repeating fixed phrases" in which word order is unchanging and "repeating collocations, in which words co-occur, although in varying order, sometimes with words intervening" (Lancashire, "Phrasal Repetends" 100). I would assert that repeating fixed phrases ("fixed phrases" hereafter), most closely represent the network of "familiar paths" that constitute a person's language. (Note that Wittgenstein repeats the sentence.) Coincidentally, cognitive neuroscience would seem to affirm this view, as illustrated above, and as Ian Lancashire notes ("Uttering and Editing" 188 ff). At present, our software deals only with repeating fixed phrases, but we have hopes of including the ability to handle repeating collocations in future releases. While the software generates lists, it has the ability to represent these lists in a way that is more easily intuited and borrows directly from the path metaphors employed above. The next phase of this talk will be an explanation of how the software works and a brief demonstration.

Practice

Locating fixed phrases by hand is a difficult task at best. Computers are far better suited to the repetitive nature of scanning through text and performing arduous comparisons. Despite their limitations, computers are capable of scanning a text for fixed phrases and displaying the results in a variety of formats, and doing so far faster than a human is capable of doing. A fixed phrase list is not relevant unless the fixed phrases are all centered about a primary word. Accordingly, our code only searches for fixed phrases centered about given primary. In order to conserve computing resources we also limit the fixed phrase search to words within a given radius of the primary. A search radius of 5 words provides enough information for most uses, though this input is not a constant. The initial implementation is straightforward. Instances of the primary in the text are located via a simple linear search. Since it is the words surrounding each instance that we are interested in, once an instance is found the phrase surrounding the primary is recorded. The search radius is used to limit the size of the phrase.

Once the text's phrases are gathered (the partial list), a full list of potential fixed phrases must be generated. Every phrase in the partial list is compared to every other phrase in the partial list. If a subset of a given pair of phrases matches, the subset is recorded. The full list is then created from the union of the partial list and the newly recorded subset phrases. Finally, the full list of potentials is filtered. Phrases that occur only once in the text are removed, while phrases that occur more than once are saved. The phrases that remain are then sorted and merged into parent/child relationships. The shortest phrases are considered parents, while phrases that are supersets of a parent phrase are declared children of the parent. This works recursively, so a child may be a parent as well. This last step makes it easier to display the fixed phrases in a graphical format. Once the phrase list is complete, it may be outputted in several formats. In addition to several varieties of formatted text, the phrase list may be drawn into a phrase list diagram. The diagram displays the phrases in a standard format, also showing the parent/child relationships described above.

This algorithm is currently implemented in both Python and Java. The Python implementation is capable of outputting the list in XML and plain-text formats, as well as in a graphical format using SVG. An external script developed for the TaPoRware project is capable of using ImageMagick to convert the SVG image into the more standardized png format. The Java implementation features a more interactive approach. A java applet presents a standard fixed phrase graph to the user. The user can then rebuild the graph around any word on the screen by clicking on it.

Works Cited

Augustine of Hippo. On Christian Doctrine. Trans. D.W. Robertson, Jr. New York: Macmillan, 1958.

Berry-Rogghe, Godelieve L.M. "The Computation of Collocations and Their Relevance in Lexical Studies." The Computer and Literary Studies. Ed. A.J. Aitken, R.W. Bailey and N. Hamilton-Smith. Edinburgh: Edinburgh University Press, 1973. 103-112.

Firth, J. R. "Modes of Meaning." Papers in Linguistics, 1934-1951. London: Oxford University Press, 1957. 190-215.

---. "A Synopsis of Linguistic Theory, 1930-55." Selected Papers of J. R. Firth 1952-59. Ed. F. R. Palmer. 1957. London, Eng.: Longmans, 1968.

Kosslyn, Stephen M, and Olivier Koenig. Wet Mind: The New Cognitive Neuroscience. 1992. New York: The Free Press, 1995.

Lancashire, Ian. "Phrasal Repetends and 'The Manciple's Prologue and Tale.'" Computer-Based Chaucer Studies. Ed. Ian Lancashire. Centre for Computing in the Humanities Working Papers, 3. Toronto: Centre for Computing in the Humanities, University of Toronto, 1993. 99-122.

---. "Uttering and Editing: Computational Text Analysis and Cognitive Studies in Authorship." Texte: Revue de Critique et de Théorie Littéraire: Texte et Informatique 13/14 (1993): 173–218.

Lyons, John. "Firth's Theory of 'Meaning.'" In Memory of J. R. Firth. Ed. C. E. Bazell, et al. London: Longmans, 1966. 288-302.

--. Linguistic Semantics: An Introduction. Cambridge, UK: Cambridge University Press, 1995.

---. Semantics. 2 vols. 1977. Cambridge, UK: Cambridge University Press, 1993.

Palmer, F. R. Semantics. 2nd ed. 1981. Cambridge, UK: Cambridge University Press, 1993.

Saussure, Ferdinand de. Course in General Linguistics. Trans. Roy Harris. Ed. Charles Bally and Albert Sechehaye. 1983. Chicago: Open Court, 1996.

Wittgenstein, Ludwig. Philosophical Investigations. Trans. G. E. M. Anscombe. 2nd ed. Oxford: Basil Blackwell, 1967.

Modelling Humanistic Activity in the Electronic Scholarly Edition

Ray Siemens, University of Victoria

A currently-accepted foundation for work in humanities computing is knowledge representation, which draws on the field of artificial intelligence and seeks to "produce models of human understanding that are tractable to computation" (Unsworth). More specifically, in activities of the computing humanist, knowledge representation manifests itself in issues related to archival representation and textual editing, high-level interpretive theory and criticism, and protocols of knowledge transfer -- all as modelled with computational techniques (McCarty).

The results of modelling the activities of the humanist, and the output of humanistic achievement, with the assistance of the computer are found in what are often considered to be the exemplary tasks associated with humanities computing: the representation of archival materials, the analysis or critical inquiry originating in those materials, and the communication of the results of these tasks (i.e. the dissemination of primary and secondary materials via electronic means). Archival representation involves the use of computer-assisted means to describe and express print-, visual-, and audio-based material in tagged and searchable electronic form (see Hockey); associated with critical methodologies that govern our representation of textual artifacts (see Machan), archival representation is chiefly textual and bibliographical in nature, often involving the reproduction of primary materials, for example, in the preparation of an electronic edition or digital facsimile of literary materials (for an exemplary initiative, see Best).

Critical inquiry involves the application of algorithmically-facilitated search, retrieval, and critical processes that, originating in humanities-based work, have been demonstrated to have application far beyond (see Lancashire, and Fortier); associated with critical theory, this area is typified by interpretive studies that assist in our intellectual and aesthetic understanding of literary works, and it involves the application (and applicability) of critical and interpretive tools and analytic algorithms on those artifacts produced through processes associated with archival representation made available via resources associated with processes of publishing and communication of results.

The communication of results involves the electronic dissemination of, and electronically-facilitated interaction about the products of, such representation and analysis as outlined above, as well as the digitisation of materials previously stored in other archival forms (see Miall); it takes place via codified professional interaction, and traditionally is held to include all contributions to a discipline-centred body of knowledge, that is, all activities that are captured in the scholarly record associated with the shared pursuits of a particular field. Knowledge in literary studies, and many other humanities disciplines, is advanced through the communication of results by interaction among those who participate in either, or both, of the processes of archival representation and critical inquiry -- more specifically, by those who do so with the intention of their resultant work to have a place in, and to take into account extant work available in, the scholarly record associated with their field.

My paper traces this model of humanistic activity, and suggests ways in which such a model has direct bearing on the design of electronic scholarly editions; in doing so, it also explores the electronic scholarly edition in the context of social theories of textual editing and textual analysis.

HyperPo: A Longitudinal Perspective

Stéfan Sinclair, McMaster University

HyperPo is a web-based text reading and analysis environment that provides many common text analysis functions, such as frequency, co-occurrence, distribution, and keyword in context lists. These functions and others are tightly integrated into the reading environment in order to promote a cyclical process of reading and analysis. HyperPo has existed since 1996 and has undergone several transformations. This presentation will outline some of the significant conceptual and architectural changes that have occurred, especially with regard to two recent developments:

1) the TAPoR portal has provided an incentive to restructure HyperPo from being a relatively autonomous online application to one that can interoperate smoothly with other tools.

2) the availability of several sophisticated open-source graphing libraries has made it possible to experiment with visualization techniques that go beyond traditional text analysis operations.

My objective is to give a good sense of where HyperPo has come from and where it is going.

Modelling the humanities scholar at work

Elaine Toms, Dalhousie University
Geoffrey Rockwell & Stéfan Sinclair, McMaster University
Ray Siemens, University of Victoria

In 2003, we conducted a survey of the research needs of humanists to better understand their working environment as well as their use of and need for computing technologies. Last fall we presented a preliminary analysis of that survey at CASTA 2003, and followed with a fuller presentation at AAL/ACH 2004. Both of these presentations provided a rich picture of the humanities world. Our respondents from a diverse group of humanities' (and social science) disciplines indicated that they used e-text for their research; about half used text analysis tools. They identified, in general, a need for readily available e-text that can be readily downloaded off the Web. They also stated some preferences concerning those texts: in a stable and legal form that is freely available from a reliable institution, with and without mark-up. In general, our respondents indicated a need for text analysis tools, although not complex tools, and expressed dissatisfaction with the tools that are currently available. While these presentations provided a rich picture of humanities scholars and their use of e-text and text analysis tools, the results were primarily descriptive.

In this next presentation, we propose to examine the inter-relationships and patterns among our data for a micro-analysis of the scholar and his/her work. We are doing the following:

1) *Modelling the relationships among our respondents by discipline, genre, form of text, geographic region, and level of collaboration*. This analysis is enabling us to answer these questions: is there a 'generic' humanities scholar in terms of their use of e-text and text tools? If there are differences (and we suspect that there are), are there 'families' of disciplines that have similar profiles? The responses to these questions relate directly to the level of customization/personalizaton that systems such as TAPoR will need to achieve to service the broadest range of clientele.

2) *Creating a task-based framework of the scholar's research environment. In our survey we asked many questions concerning the work that the scholar does, the tools they use, how they use those tools in their research, and the tools/techniques they would like to use*. From these questions we have identified to date the frequency of use of e-text and tools. In the current analysis, we have adopted a task analysis methodology borrowed from human computer interaction to define a task framework for the humanities scholar. Such a technique is normally employed at the requirements stage of software development. In this case, we are using it to build a profile of the scholar's working environment using a scholar-centred - a classic user-centric - approach to identifying research needs and requirements for the scholar at work.

These in depth analyses of our survey data are intended to model the humanities scholar at work. Examining the differences in respondent profiles will indicate the types of humanities scholars that exist, and suggest user models. Understanding the relationship between scholar's needs and tool use/preferences will enrich those user models. The models have the potential to aid in identifying systems requirements for the humanities.

Forms of Attention: Digital Humanities Beyond Representation

John Unsworth, the Graduate School of Library and Information Science, University of Illinois

Humanities research computing in the 1990s was focused on editing, modeling, and other forms of representing primary resources. This activity raises some vexing problems in the area of copyright, intellectual property, and contract law, and there is no sign that these problems are getting any less vexing in the 21st century. What are the possibilities for research computing in the humanities that is not dependent on the representation or republication of primary resources? This paper will explore that question.

The Text of the database : can you read me ?

Christian Vandendorpe, Université d'Ottawa

We commonly use the term "navigate" to describe the activity of the "internaut" (a term reminiscent of the famous act of navigation performed by the Argonauts in ancient Greece). It is interesting to note that the same navigational metaphor has been used in the past to describe the activity of reading a book. In *Paradiso* 2, Dante compares his book to a large ship that the readers can follow in a small boat, but not in the open sea. This was a way of establishing a clear hierarchy between writing and reading: the reader was just following, from afar.

On a computer screen, however, the reader is expected to take a much more active role in the navigation. In fact, in the digital world, both the nature of text and the nature of reading are undergoing a profound transformation, enriching the palette of that complex ability called "reading".

The representation of the act of "reading" was traditionally one of a person profoundly absorbed in a book: be it an old man, a monk or a saint, be it the Virgin Mary (most of the time represented with a book in her hand), be it an old woman like Rembrandt's mother. And the ideal type of book is the repository of a long narrative in which the reader would be immersed for many hours during many days.

A new conception of reading began to emerge at the 18th century, according to the historians of the book: an extensive conception of reading replaced the intensive one which dominated before. That extensive model gained even more in importance with the success of the newspaper in the 19th century, the expansion of the magazine in the 20th century and, more recently, the advent of the Internet. The “extensive” reader does not absorb the text word by word, but scans rapidly through the titles and the various bits of text until he or she finds something that seems worth being read, at least for a few paragraphs.

Besides these two main ways of reading, the intensive and the extensive, there is a third one, which involves the pursuit of knowledge through the interrogation of texts. The inquisitive reader who embarks on this kind of activity is no longer a somewhat passive entity expecting to be taken on a trip by a text or a series of texts. He or she is scanning a multitude of texts in order to test a hypothesis. For this kind of reader, the text may be compared to an open gold mine, in which a suitable nugget of knowledge is expected to be found through the treatment of millions of words.

That method of reading is not entirely new. In fact, one can suggest that it was already somewhat practised in the 12th century when the first indices made their apparition in manuscripts. But it was a very marginal activity, compared to the more common form of reading. This practice, however, gained more recognition in the 18th century with the first encyclopaedias. Today, since the computer is part of our environment, reading to find an answer to a question is becoming a much more important part of our daily reading activities. Sure, we still occasionally read novels in a linear fashion, and we read magazines in a browsing fashion, either on paper or on the Internet. But the success of Google confirms that inquisitive reading is a very big part of our everyday experience.

The database is thus fast becoming a new repository of text. This is true not only for general search engines, but also for specialized content. For people engaged in research, the database is the best way to make vast amounts of data easily searchable, whatever the discipline.

As an illustration, I shall present the database for the study of dreams that we have been building, thanks to a grant from the SSHRC, since 2001.

The aim of this database is to provide the data and tools for an investigation of the dream narrative in many different cultures. More specifically, our conviction is that the database will: (a) illuminate the relationship between dream narratives and the society in which they are told, offering contrasting characteristics between different cultures and different time frames; (b) show the evolution of the interpretation of dreams; (c) permit a comparative study of literary dream narratives and everyday dream reports.

Let it be clear that a database is not an anthology. The anthology is a collection of the most sublime accomplishments of a genre. The reader of an anthology is expected to contemplate the various “flowers” collected by a specialist in order to get a glimpse of the best manifestations of a genre. He is like an amateur savouring various samples of wine or cheese that have been chosen by an expert.

With our database, we are not so much preoccupied by the exemplary nature of the texts collected as by offering as complete a collection as possible of the various occurrences of dream narratives, particularly those present in literature. Given the nature of the narrative, and particularly of the dream narrative, we have decided to include a wide range of texts, from the short notation of a dream to much more elaborate narratives, and to cover a great variety of genres: the novel, theatre, poetry, diaries and even books on dream interpretation.

In a first phase, from 2001 to 2004, we have explored the field of French literature from the Middle Ages to today. Now we are extending our search to Antiquity (Greek and Latin literature, the Bible), and to the other great literatures of the world (English, German, etc.) We have already more than 800 texts from 200 authors. We plan to double this number within the next three years.

Now let us consider the design of the database and its interface.

A database should not be simply a repository of hidden data that only the knowledgeable user can summon to light – a concept that is still much too frequent on the Web. There are too many opaque databases, even when they are built for public sites and intended theoretically to fulfill an educational mandate. For example, the virtual museum of Canada (www.virtualmuseum.ca/) offers on its homepage various choices, among which a "Virtual Gallery". As a new visitor, you might think that by clicking on that link you would find all the paintings available in the database. In actual fact, if you follow that link, you will only see a small sample of the contents of the Gallery. In order to see the works of artists who are not on that list, you have to type their name in a box. The problem is: "How do you find an artist whose name you have never heard?" As a result, if you don't know the names of John Varley, Janet Logan or Jean-Paul Riopelle, you won't be able to find them or learn anything about them by visiting this so-called virtual museum. This kind of opaque database is only useful for those who already know what they are looking for.

That model has no educational value and is the exact opposite of what a database should be. A database open to the public should not be conceived as a "black box" from which you extract data through queries. On the contrary, it must be made as readable as possible, in the form of a space where the user will be able to browse easily through the various records.

How do we achieve that goal with www.reves.ca?

First, all the data can be accessed by the reader, either sequentially, or randomly. There are no restrictions.

Second, the reader can also access the texts through a list of all the authors present in the database.

Third, an elaborate search engine allows the reader to find all the records that contain a certain word, a series of words, and/or all the records that meet certain conditions: country, era, type of text, characteristics of the narrative, types of characters appearing in the dream, and so on.

When you arrive on the page displaying the text of a dream, you will find, naturally, the name of the author, the date and title of the original publication. But there are also a lot of editorial marks. A short paragraph in the margin presents the context of the work in which the dream appeared and, if necessary, a few notes explaining obscure details in the text. Moreover, each dream is identified by a title and a subtitle, which are designed to guide the reader and awaken their interest. Occasionally, a hyperlink ("Interpretations") will point to a smaller window displaying various interpretations that have arisen from the text, as is the case with the famous dreams of Descartes. Each narrative is also accompanied by details on the type of work where the dream appeared ("Type de texte") and the type of narrative used ("Forme"). A collection of keywords describe the psychological content ("Mots-clefs"), another describe the characters ("Personnages"), another the category of the dream ("Type") and yet another the main themes ("Thème"). All these editorial marks multiply the chances that the readers will find meaningful relationships satisfying their research needs or their curiosity.

Finally, since the database aims to foster research on the dream experience, we have added the possibility for the user to create their own writing space. By clicking on the link "Ouvrir une session", any user is able to make annotations on each record, and then save them and retrieve them later. These annotations will also appear in the window of the appropriate dream narrative when the user comes back and browses through the database, just as you would be able to read again the notes you wrote in the margins of a book. For the moment, all those annotations are private and visible only to the user who has created them, but a new function will soon allow the users to make them public if that is their wish. In that case, the record of a dream narrative will be accompanied by the names or pseudonyms of the various searchers who have commented on it, and other readers will be able to access these comments, thus adding more meaning to the reading experience and helping to create a truly interactive space for the exploration of dreams.

By giving a new dimension to reading and writing, these tools are transforming our relationship with texts. The database becomes an object in which readers can browse, read in a linear fashion or scan and interrogate in order to expand their understanding of the complexity of life.

Whose funeral? A case study of computational methods and reasons for their use or neglect in English literature.

Claire Warwick, School of Library, Archive and Information Studies, University College, London

This proposal considers whether there is evidence that scholars in the field of English literature are actually using computational techniques, such as text analysis, authorship studies, the creation of critical editions, or even the use of online resources or electronic journals. Research was conducted by examining publications in a selection of journals from 1990 to 2003. The sample included both specialist humanities computing journals, *Literary and Linguistic Computing and Computers and the Humanities,* an electronic-only journal *Early Modern Literary Studies,* and more traditional printed journals., *The Yearbook of English Studies, Review of English Studies, Prose studies,* and *PMLA.*

Findings:

The number of articles in humanities computing journals has declined slightly compared to the early 1990s, to an average of about five a year. Yet this was relatively high compared to journals in English literature. Clusters of articles occasionally appeared in special issues on computational methods. Perhaps predictably, most articles in English literature journals were found in *EMLS*. Yet the majority of them simply cited online resources or e-journals. Even to find a URL cited in *PMLA* is a rarity. Most worryingly a few articles, especially in *RES*, seemed to have used electronic resources or computational methods yet did not describe their methods or cite sources. The few articles that used computational methods all discussed one subject: the controversy over Don Foster's attribution of *A Funeral Elegy* to Shakespeare.

Discussion:

Why should digital resources and computational techniques have been so neglected? The research considers factors such as a lack of knowledge and technical support, or resources of the right kind. More complex factors include critical trends that have meant that textual analysis of any kind has been relatively neglected in favour of historical and cultural criticism. It also discusses the perceived prestige of digital resources, and speculates about why the use of online resources may not be acknowledged.

The *Funeral Elegy* controversy provides a case study of circumstances in which the use of computational techniques was noticed and adopted by mainstream scholars. The paper argues that a complex mixture of a canonical author (Shakespeare) and a star scholar (Foster) brought the issue to prominence, by making clear how and why computational methods were used. Authorship studies is also an area that can be researched by both traditional historicist scholars and those who use computers.

Conclusion.

The *Funeral Elegy* debate shows that if the right tools for textual analysis are available, and the need for, and use of, them is explained, some mainstream scholars may adopt them. Despite the current emphasis on historical and cultural criticism, scholars will surely return in time to detailed analysis of the literary text. Therefore researchers who use computational techniques must publish their result in literary journals as well as those for humanities computing specialists. We must also realise that the culture of academic disciplines is relatively slow to change, and must engage with those who use traditional methods. Only when all these factors are understood, and are working in concert, may computational analysis techniques truly be more widely adopted.

Testing TAPoR with Medieval French Poetry

Jenna Wells & Madeleine Jeay, Department of French, McMaster University

What tools are necessary to make the electronic and computational study of medieval texts a reality? What new readings of old texts can be generated using computational tools? What does our experience today tell us about the types of tools available? What tools should be developed in the future?

This paper will discuss why conversion to electronic format is of particular importance for medievalists. We will focus on the specific needs that medieval electronic text research demands and the tools necessary to meet those needs. The specific corpus that we will use to exemplify our discussion is a sub-part of the *Eustache Deschamps Enumerative Poetry Corpus*, which concerns lists of allegories in ballads and the networked relationships between those allegories across texts. This corpus is part of the larger *Hyperliste* on-line textual database, housed at McMaster University under the responsibility of Dr. Madeleine Jeay.

The *Hyperliste* project is funded by the SSHRC and has three main objectives: (a) to create an electronic textual corpus; (b) to create a new, hypertextual mode of publication for medieval lists in literature; (c) to create a database of special glossaries.

The *Hyperliste* project offers a critically annotated electronic edition of all the literary lists that have been preserved from the XIIIth to the XVIth centuries. It includes poems, dramatic monologues and joyous sermons. The XML format allowed us to demonstrate the hypertextual, non-linear, circulation of topic from one text to another. This makes it easy for the reader or researcher to read the texts in a non-linear fashion, a feat that is difficult to achieve when lists are in analog (paper) format only.

Hyperliste contains a series of inventories which represent terminological resources that can serve as a key to a period where unilingual vernacular dictionaries had not yet been developed. *Hyperliste* will be a tool for the creation of special glossaries specific to the different aspects of everyday life found in lists concerning, for example, utensils, tools, professions, food, illnesses, injuries, etc.

An attempt was made to analyze the allegory corpus using TaPOR and the tools contained within it. Part of the *Deschamps Allegory Research Project* was to test TaPOR's ability to handle the analysis of discontinuous texts. This paper will address current limitations of the TaPOR system and toolset encountered during our research. We will also present possible improvements and additions that could be made to TaPOR to make it more useful for electronic medieval research.

Sponsor's Message

Open Sky Solutions

Open Sky Solutions is pleased to support the CaSTA Face of Text conference. Our work with McMaster and the TAPoR group across Canada has been extremely rewarding and fruitful, and we look forward to a similar experience at the Face of Text conference.

Open Sky Solutions has been working closely with the Humanities Computing group at McMaster University since March of this year, developing the online text analysis portal for the TAPoR group. We have adapted industry standard open source components to produce a highly functional portal that incorporates innovative ideas to an extent not possible with a commercial portal solution, and further, remains open for ongoing growth. We believe the TAPoR portal successfully realizes the promise of open source software. We are also very proud of the partnership that we have built with TAPoR. Our development process is highly iterative and interactive, and requires a significant commitment from McMaster. That commitment has produced, in a short period of time, an extremely valuable contribution to the text analysis community.

Dr. Rockwell will demonstrate the TAPoR portal during the opening session on Friday in the McMaster Council Chambers. Open Sky Solutions is very proud of the portal and we look forward to hearing your reactions.

We hope you enjoy the conference and your stay in Hamilton.

Open Sky Solutions (www.openskysolutions.ca) is a consulting and development company, focused on applying and adapting open source software for companies developing enterprise solutions. We implement innovative software solutions built around industry standard open source components, using a highly iterative development methodology to produce exceptional functionality and ensure ongoing maintainability and growth. Open Sky Solutions combines expertise in object-oriented analysis and design with core competencies in open source software to efficiently deliver robust software solutions.

Presentation Schedule

Friday, Nov. 19	
10:00 - 11:00am	**Jerome McGann**
11:30 - 1:00pm Session 1:	1. **David L. Hoover** **2. Ray Siemens** 3. **Patrick Juola & John Sofko**
2:00 – 3:00pm	**John Unsworth**
3:30 - 5:30pm Session 2:	1. **Gary W. Shawver & Oliver Kennedy** 2. **Susan Brown** 3. **Michael Best** 4. **Mikaël Roussillon, Bradford G. Nickerson, Stephen Green & William A. Woods** 5. **Dominic Forrest** 6. **Christian Vandendorpe**
Saturday, Nov. 20	
8:30 - 9:30am	**Julia Flanders**
10:00 - 11:30am Session 3:	1. **Claire Warwick** **2. Paul A. Fortier** 3. **Paul Scifleet & Concepcion S. Wilson**
12:30 - 1:30pm	**John Bradley**
2:00 - 3:30pm McMaster Presentations	1. **Jenna Wells & Madeleine Jeay** 2. **William Coleman & Andrew Mactavish** 3. **Nicholas Griffin & James Chartrand** 4. **Stéfan Sinclair**
4:00-5:30pm Session 4:	1. **Atefeh Farzindar & Guy Lapalme** 2. **Tobias Kalledat** 3. **Stan Ruecker & Zachary Devereux**
Sunday, November 21	
9:00-10:00am	**Jean-Guy Meunier**
10:30 - 12:30pm Session 5:	1. **Pamela Asquith & Peter Ryan** **2. Eugene W. Lyman** 3. **Marc Plamondon** 4. **Jason Boyd** 5. **Elaine Toms, Geoffrey Rockwell, Stéfan Sinclair & Ray Siemens**
12:30 – 1:30pm	**Stephen Ramsay**
2:30 – 3:00pm	**Closing Remarks**